Indoor and Outdoor Location Estimation in Large Areas Using Received Signal Strength

by

Kejiong Li

A thesis submitted to the University of London in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Electronic Engineering and Computer Science
Queen Mary, University of London
United Kingdom

July 2013

To My Family

Acknowledgments

First and foremost, my deepest gratitude goes to Dr John Bigham, my supervisor, for his consistent support, patient guidance and professional supervision. I deeply appreciate all his contributions of time and work to make my PhD experience productive and stimulating. His knowledge and enthusiasm towards research have inspired me, and encouraged me all the time.

My heartfelt thank goes out to Dr Eliane L Bodanese for her guidance and support that have given me. Her valuable ideas and helpful advice have helped immensely in writing this thesis. My sincere appreciation also goes to Dr Laurissa Tokarchuk for giving me the opportunity to work in EPSRC ImpactQM KTA project. I am very grateful to Dr Karen Shoop and Dr Paula Fonseca who have provided me to work as a teaching assistant throughout my studies.

I would like to address special thanks to Dr Peng Jiang. He has generously given me support and professional advice especially at my most difficult time. I also wish to thank Dr Gareth Tyson for making a number of helpful suggestions regarding the final version of this thesis.

I am extremely grateful to my colleagues in Queen Mary. The group has been a source of friendship, good advice and collaboration.

Finally, I would like to thank my beloved family. For my parents who raised me, love me and support me for years. And for my dear husband, Mr Chao Ning, whose patience and faithful support are so appreciated.

Abstract

Location estimation when deployed on wireless networks supports a range of services including user tracking and monitoring, health care support and push and pull marketing. The main subject of this thesis is improving indoor and outdoor location estimation accuracy using received signal strength (RSS) from neighbouring base stations (BSs) or access points (APs), without using the global positioning system (GPS) or triangulation methods. For the outdoor environment, state-of-the-art deterministic and probabilistic algorithms are adapted to exploit principal components (PCs) and clustering. The accuracy is compared with K-nearest neighbour (KNN) algorithms using different partitioning models. The proposed scheme clusters the RSS tuples based on deviations from an estimated RSS attenuation model and then transforms the raw RSS in each cluster into new uncorrelated dimensions, using PCs. As well as simple global dimensionality reduction using PCs, the data reduction and rotation within each cluster improves estimation accuracy because a) each cluster can model the different local RSS distributions and b) it efficiently preserves the RSS correlations that are observed (some of which are substantial) in local regions and which independence approximations ignore. Different simulated and real environments are used for the comparisons. Experimental results show that positioning accuracy is significantly improved and fewer training samples are needed compared with traditional methods. Furthermore, a technique to adjust RSS data so that radio maps collected in different environmental conditions can be used together to enhance accuracy is also demonstrated. Additionally, in the radio coverage domain, a non-parametric probability approach is used for the radio reliability estimation and a semi-supervised learning model is proposed for the monitoring model training and evolution according to real-time mobile users' RSS feedback.

For the indoor environment, an approach for a large multi-story indoor location estima-

tion using clustering and rank order matching is described. The accuracies using WiFi RSS alone, cellular GSM RSS alone and integrated WiFi and GSM RSS are presented. The methods were tested on real indoor environments. A hierarchical clustering method is used to partition the RSS space, where a cluster is defined as a set of mobile users who share exactly the same strongest RSS ranking set of transmitters. The experimental results show that while integrating of WiFi RSS with GSM RSS creates a marginal improvement, the GSM data can be used to ameliorate the loss of accuracy when APs fail.

Contents

List of Abbreviations

AGPS	Assisted Global Positioning System
AMISE	Asymptotic Mean Integrated Square Error
AOA	Angle of Arrival
AP	Access Point
BS	Base Station
COTS	Commercial-of-the-shelf
EE	Electronic Engineering
eNB	evolved Node B
GPS	Global Positioning System
KDE	Kernel Density Estimator
KL-distance	Kullback-Leibler distance
KNN	K-Nearest-Neighbour
KNN-VPM	K-Nearest Neighbour-Venn Probability Machine
KS-test	Kolmogorov-Smirnov test
LOS	Line-of-sight
LTE	Long Term Evolution
MAP	Maximum A Posteriori
ML	Maximum Likelihood
MIMO	Multiple Input Multiple Output
MMSE	Minimum Mean Square Error
MS	Mobile Station

MUSIC	MUtiple Signal Classification
NLOS	Non-line-of-sight
NN	Nearest-Neighbour
PC	Principal Component
PCA	Principal Component Analysis
PCA-Intersection	Intersection after Principal Component Analysis
PCA-KDE	Kernel Density Estimator after Principal Component Analysis
PCA-WKNN	Weighted K-Nearest Neighbour after Principal Component Analysis
pdf	probability density function
QoS	Quality of Service
RF	Radio Frequency
RFID	Radio Frequency Identification
RMSE	Root Mean Square Error
RS	Relay Station
RSS	Received Signal Strength
RTT	Round Trip Time
SON	Self-Organizing Network
SISO	Single Input Single Output
TDOA	Time Difference of Arrival
TOA	Time of Arrival
TOF	Time of Flight
TTFF	Time to First Fix
VPM	Venn Probability Machine
WKNN	Weighted K-Nearest-Neighbour

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

The proliferation of wireless technology and mobile computing devices has fostered the expanding demands and unpredictable traffic demands across a variety of Internet services. To achieve a high data rate and seamless coverage, the complexity of radio access networks continues to increase, and the expenditure on operational tasks, such as network planning, deployment and network optimisation, is rising to an unprecedented level. For such a complex situation, it is perhaps too difficult to meet users' demands through conventional cellular networks which require a large amount of manual labour and have limited capacity. Self-Organizing Network (SON) has been considered to be an effective way to tackle these challenges supported by Long Term Evolution (LTE) networks in 3GPP standards [4]. SON has the potential to support the integration of network planning, configuration and optimisation into a single and mostly automated process, requiring minimum human interaction and deployment effort [5]. It is proposed for high speed wide-area wireless networks, where many advantages are expected in terms of coverage, throughput and Quality of Service (QoS) provisioning [1] [5].

Network service providers want to apply SON approach into realistic wireless net-

works, so that they can not only reduce the costs associated with human operational involvement, but also optimise network capacity, coverage, performance and adaptability in the presence of a variable network environment. To provide adequate radio coverage with minimum cost for mobile users in a constantly changing network configuration, it is essential to have the ability to predict radio propagation and traffic demand behaviour accurately. These can depend on topographical features and network configuration. However, the constraints of network capacity, hard-to-predict mobile station (MS) movements, the complexity of realistic propagation environment (e.g. multipath and shadowing), the manpower and equipment costs for coverage measurement, all present significant challenges to efficient radio resource allocation. SON functions give a promising opportunity for automated localised, distributed as well as centralised functions to accomplish wireless access network planning and optimisation in the context of a variable real environment.

This thesis investigates the use of signal strength received by MSs for positioning and radio coverage prediction, in order to monitor the received signal strength (RSS) distribution in a real-time and radio coverage status. This gives the fundamental support for the self-processes (configuration, optimisation) in SON. The proposed approach could also be used to assess the network state and affected civilians in emergency or disaster situations.

(1) Potential Usage: Run Time Measurement for SON

In SON, the use cases are mainly divided into three procedures: self-configuration, self-optimisation and self-healing. No matter which category, the network measurement phase plays a very important role in the process of SON, as illustrated in Figure 1.1. In this figure, the red arrow means the flow line which denotes the logical flow in this procedure and the blue "splodge" illustrates a cell or a site is in a failure state. During the measurement phase, the measured data, e.g. user' RSS feedback, mobility, traffic patterns and interference are collected through information exchange among base stations (BSs) or relay stations (RSs) in order to analyse and assess the network behaviours, and

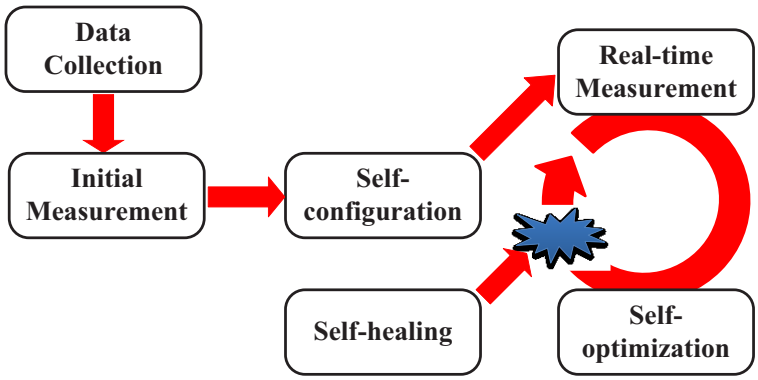


Figure 1.1: The operational process applied in self-organizing networks

trigger adequate actions. The self-configuration process is designed for newly deployed BSs to automatically monitor for bad coverage or shortage of offered bandwidth. It helps to establish whether there is need for a new BS on a site and to determine the optimal location for this new BS. In the self-optimisation phase, the processed measurements are periodically used to monitor the real environment, and (self-) adjust the operational algorithms and parameters in response to the changing conditions of the environment. If a cell or a site is in outage or in a failure state, self-healing techniques will take some recovery actions to ensure better radio coverage in these failure areas by adjusting network parameters and algorithms in nearby BSs (e.g. antenna tilting, sector power changes). Thus, the measurement phase is crucial in all the procedures in SON as it can provide important information to the different configuration and optimisation tasks that manage the response to the changing conditions of the environment.

(2) Location Estimation

Location estimation has been a hot topic in the past years. The location estimation systems are designed to estimate the position of a MS. Outdoor location estimation has been possible for a long time using global positioning system (GPS). Other techniques are offered by e.g. Google, when the satellite is not in direct line of sight, but these are very inaccurate as they are often simply based on base station locations. GPS drains the phone battery and for this reason many people do not turn the function on. For

example, you may not want to drain your battery at a music festival because you will want to make calls. Also, in the campus, you are often moving in and out of buildings and GPS does not work well indoors. Likewise, indoors location estimation based on the readily-collected received signal strength (without the use of special Radio Frequency Identification (RFID) tags), is also inaccurate. Thus, the issue on how to give high precision estimates both indoors and outdoors without using the computational and battery resources of GPS to allow easy application development is one of the important challenges in this thesis.

This thesis concentrates on building models to estimate a user's location in outdoor and indoor environments. The models are constructed from observed coverage patterns in known antenna configurations and the accuracy of the predictions based on these models in different scenarios is evaluated. The models are created by the analysis of RSS data collected from historical data, i.e. during a training phase, and monitoring of current mobile users' RSS feedback to assess the current RSS distribution.

In an outdoor environment, the RSS data collected during the training phase is partitioned into relatively homogeneous regions with respect to RSS, by using a clustering tool. This partitioning of the RSS space is used to support location estimation that can be accurate enough to support resource allocation in SONs, without using the computational resources of triangulation. Five issues will be taken into account in outdoor localisation: the first issue is how to effectively quantify the importance of each transmitter and choose the most reliable transmitters for different regions; the second issue is how to create the number of clusters that can represent the features of the geographical patterns in the area of interest; the third issue is how to allow for the fact that different correlations[1] between BSs or access points (APs) signals occur at different points in RSS in a cluster area; the fourth issue is how to reduce the useful information into relatively lower dimensions by a suitable transformation in each cluster while reducing the computational complexity and data requirements given the large scale of the area; the

[1]The term "correlation" refers to a process for establishing whether or not relationships exist between two variables.

last one is how to effectively accommodate the variation of signal strength over different environmental conditions without having to rebuild the radio map repeatedly. Different proposed estimation methods for outdoor location approximation based on clustering models are tested for their estimated values in both data from a simulated propagation model and a network planning tool, and data collected from the real environments. The performances of estimation accuracy are evaluated.

In many potential applications approximate location information such as the room they are in, or the room segment if the area is large, is adequate. The objective of indoor localisation in this thesis is to locate mobile users in a specific room in a large multi-storey environment where both GSM RSS and WiFi RSS are being used. In addition, the issue of how to quickly locate mobile users in emergency events at a Shopping Mall is also taken into account. The event is an incident that requires the fast evacuation of the mall where some of the exits may be blocked or congested (a warning prior to the incident), or post-incident where people are stuck in the mall and not sure what to do or want help. The proposed methods are tested on two-floor of the Electronic Engineering (EE) building in Queen Mary campus and the lower ground, ground and first floors of the Stratford Westfield Shopping mall in London, which is claimed to be the largest of its kind in Europe.

(3) Coverage Prediction

The clustering partitioning approach being developed also supports coverage prediction. Previous research on cooperative control for radio coverage involving the physical layer [6] [7] have shown that cooperative control is a novel way to handle heterogeneous traffic and to improve the whole network performance significantly. The main idea of cooperative control is to use the radio frequency (RF) domain optimisation to increase utilization of the limited frequency spectrum at reasonable costs. In other words, cooperative control could cut the cost and time spent on network deployment by optimising the RF domain according to perceived traffic distribution and propagation environment. In order to reason about the best cooperative coverage in novel situations, it is critical to

propose and develop an approach that can build models of the expected radio propagation that approximate to the real world in the area of interest. Cooperative control does not necessarily involve only adjustments to the RF coverage; it can be combined with, for example, dynamic frequency allocation. However, any changes to the RF coverage require an assessment of the impact on the traffic in the area. This is a difficult problem in a real environment and also is the second challenge considered in this thesis.

To use the users' RSS feedback for the radio coverage estimation, previous researchers need to obtain the users' exact location information. However, in the proposed approach, the target area is divided into different clusters. One of the advantages in this approach is the exact location is not required, the location information is cluster based. For each cluster, there is a probability density model to represent the probability density function (pdf) of RSS from one antenna system at a specific antenna configuration. This information can be used to maintain service reliability in the given region of a wireless network. This is one possible use of this research.

1.2 Research Objective

This thesis is interested primarily in the problems of estimating a mobile user's location indoors and outdoors based on the RSS and secondly in the prediction of outdoor radio coverage. The important observable in this thesis is RSS.

Why use the RSS for localisation?

In the early years, location applications employed only a small number of sophisticated receivers or antenna arrays, typically as few as two or three, particularly for long-distance positioning. These positioning systems preferred to use triangulation techniques, e.g. Time of Arrival (TOA), Time Difference of Arrival (TDOA) and Angle of Arrival (AOA) over RSS to achieve high location accuracy. However, the recent proliferation of wireless devices and networks has enabled a larger number of observation

points, e.g. APs that are relatively close to the mobile target. Furthermore, the new wireless applications and services for which the RSS approach is suitable have been developed, particularly in indoor and urban non-line-of-sight (NLOS) environments. Hence, an RSS-based approach is a potentially viable, cost-effective solution that can be applied to a broad range of applications while providing comparable location accuracy.

Despite lower location accuracy than the time-based techniques, the RSS-based localisation is a simple, low-complexity method that can be integrated into another type of location system for a hybrid approach. Particularly, RSS values are readily available in most wireless systems without additional hardware or system modifications. In fact, RSS information is required by many wireless standards and specifications for the purpose of basic radio functions such as clear channel assessment, link quality estimation, handover, and resource management [8].

How to use RSS for Localisation?

Performing RSS-based localisation is a challenging task due to multipath effects in outdoor and indoor settings. These effects include shadowing and reflection. Thus, the RSS measurements will be attenuated in unpredictable ways due to these effects. To tackle these challenges, this thesis investigates how to improve indoor and outdoor location estimation accuracy using received signal strength from neighbouring BSs or APs, without using the GPS or triangulation methods. The main contribution of this thesis are summarised as follow:

- In outdoor environments, the aim is to partition the target area into different regions where different localisation algorithms can be used to build a model. A clustering scheme that uses the Affinity Propagation method is developed that creates clusters based on deviation RSS feedback from mobile users who are at different physical locations, and utilises the Venn Probability Machine (VPM) algorithm to predict the probability of cluster membership. It manages the trade-off between estimation accuracy of cluster identification and the number of clusters to

select a better clustering scheme.

- To improve outdoor location estimation accuracy using the RSS from neighbouring BSs without using GPS, state-of-the-art deterministic and probabilistic algorithms are adapted to exploit principal components (PCs) based on the proposed clustering scheme. There is a high correlation between the RSS from different BSs in a real environment, which leads to a decrease in estimation accuracy unless accounted for. Therefore, the aim is to estimate user location despite the potentially high correlation RSS values. In each cluster, Principal Component Analysis (PCA) is used to transform the RSS into uncorrelated RSS tuples. The novel deterministic algorithm, called Intersection after Principal Component Analysis (PCA-Intersection), aims to find the most likely intersection area of more than three BSs circles (coverage areas) in geographical spaces to calculate a mobile user's position after RSS transformation. A Kernel Density Estimator after Principal Component Analysis (PCA-KDE) belongs to the probabilistic category. This algorithm estimates the MS location by using the KDE to build an adjusted RSS probability distribution. Additionally, a technique to adjust the RSS data without rebuilding the radio maps collected in different environmental conditions is developed. This can also be used to enhance accuracy.

- To support real-time monitoring of radio coverage in a dynamic environment, a semi-supervised learning mechanism is described. Radio coverage probability models based on RSS from BSs in an outdoor environment are created. The proposed clustering is also used to partition the RSS space and a nonparametric probability approach is used to reliably estimate the radio coverage in each cluster and also test for discrepancies in the RSS coverage that may occur over time. It is assumed that data can be collected periodically from the physical environment. The analysis of discrepancies is based on models constructed from historical data and monitoring of current RSS from the MSs.

- An indoor positioning algorithm using clustering and ranking patterns is proposed

for locating mobile users in a large scale multi-story indoor environment where only GSM RSS data, or only WiFi RSS data, or hybrid RSS (e.g. GSM RSS and WiFi RSS) are collected from mobile phones. Moreover, this approach is further improved to use GSM data only to ameliorate the loss of accuracy when APs fail.

1.3 Thesis Structure

The remainder of this thesis is organised as follows:

Chapter 2 presents the basic concepts of network planning and radio propagation models, and then introduces a comparison of the location determination systems for typical wireless location technologies. Finally, it reviews the literature in coverage prediction.

The basic concepts and methods of location fingerprinting are presented in **chapter 3**. The emphasis is on the mathematical formulation and structuring the methods according to their theoretical background. The methods covered in **chapter 3** can be divided into deterministic and probabilistic approaches.

In **chapter 4**, the proposed outdoor positioning mechanism, its corresponding transmitter selection method and the clustering scheme are described.

The PCA-Intersection method is described in **chapter 5** and its performance is evaluated by comparing it with common existing deterministic algorithms using simulated and real data.

The PCA-KDE method is proposed in **chapter 6** to estimate the MS location by using KDE to build RSS distribution in each cluster.

Chapter 7 illustrates the proposed method to adjust RSS in different environmental conditions in order to improve the estimation accuracy in dynamic environment.

Chapter 8 concentrates on building models to predict radio coverage probability that are based on clustering (RSS from BSs). For each RSS partition, a nonparametric probability approach is used for the radio coverage reliability estimation, and the analysis is based on models constructed from historical data and the monitoring of the RSS from the MSs.

In **chapter 9**, the approach for indoor location estimation that integrates RSS data from both WiFi and GSM networks is presented.

Finally, the last chapter concludes the thesis and some suggestions are made as to how the work could be extended.

Chapter 2

Network Radio Characteristics and Localisation Methods

2.1 Introduction

This chapter presents an introduction of the essential principles of network planning in section 2.2. Section 2.3 briefly describes the propagation characteristics and classical empirical radio propagation models. Section 2.4 first provides an overview of the main positioning techniques: Time of Arrival (TOA), Time Difference of Arrival (TDOA), Angle of Arrival (AOA) and received signal strength (RSS), and then compares these four positioning approaches on the basis of described performance parameters. Section 2.5 reviews some typical existing positioning systems and the technique aspect of radio coverage prediction is introduced in section 2.6.

2.2 Radio Network Planning

The process of network planning aims to satisfy the ever-increasing demand for network coverage and capacity from network vendors and mobile users. Among the whole network

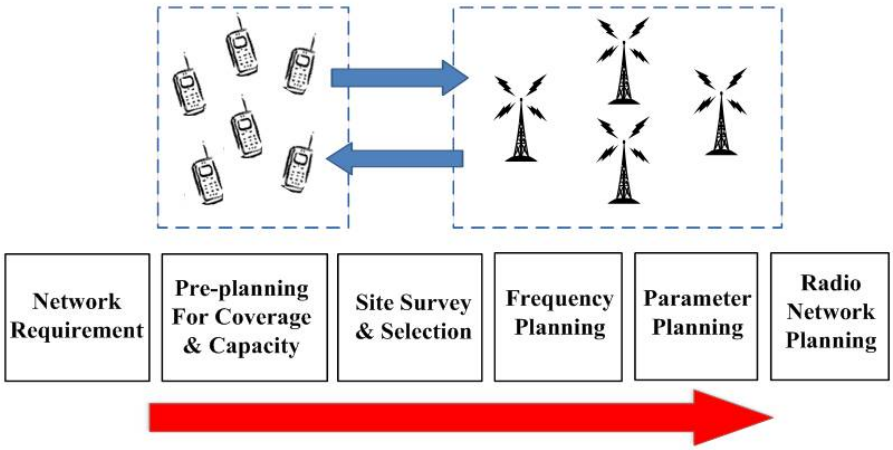


Figure 2.1: The radio network and network planning process [1] [2]

design process, radio network planning might be the most important stage, and its results affect the performance of a wireless network. The radio network is a part of the wireless network, including the BSs, MSs and the interface between them [1] [2]. In the radio network, BSs have the capability of communicating with MSs within a certain coverage area, and of maintaining network quality. The process of radio network planning includes five main phases before the final plan is generated, as shown in Figure 2.1. In this figure, the square shape represents the steps in this process and the red arrow shows the sequence of these steps. The blue arrows show the communications between BSs and mobile users in wireless networks.

Radio network planning starts with network requirement phase, which focuses on estimating and collecting all relevant parameters (e.g. potential traffic demands, service type provision, and related signal strength) of the considered area. Then, determining the coverage and capacity criteria according to the data collected in the former stage. After the pre-planning phase, the site search process starts. The process of site selection and site survey is to identify specific areas for prospective sites based on the coverage and capacity plans, and to generate the reports about information of the candidate sites. When the best sites are chosen, making an accurate frequency allocation plan could not only diminish the effect of interference, but also maintain the desired network quality. Finally, the objective of parameter planning is to pre-define and optimise the parameters

related to signalling, radio resource management, and handover, etc. so that those data can help to provide seamless communication with minimal interference [2].

2.3 Radio Propagation Model

It is critical for wireless communication systems to have the ability to predict accurately the radio propagation behaviour. Because conducting site measurements are not only expensive but also time consuming, it is valuable to develop a low cost, convenient alternative that is able to determine optimum BS location, estimate coverage, etc. Hence, it is essential to develop a proper radio propagation model that can reflect the current system propagation characteristics and estimate the signal characteristics effectively for the purpose of wireless network planning during preliminary deployment.

The radio propagation model describes the signal attenuation from the transmitter to the receiver antenna as a function of distance, carrier frequency, antenna heights and other significant parameters like terrain profile (e.g. urban, suburban and rural). Based on the radio propagation model, a wireless radio channel has three main characteristics: path loss, shadowing (slow fading) and fast fading [9].

- **Path Loss** is caused by the dissipation of power radiated by the transmitter. It determines how the average received signal power falls off relative to the distance between the transmitter and the receiver.
- **Shadowing** is a kind of fading caused by larger movement of a mobile, or by an obstruction, like a hill or large building that obscures the main signal path between the transmitter and the receiver. It is often modelled as a log-normal distribution with a standard deviation according to the Log distance path loss model. In a slow fading channel, the channel impulse response changes much slower than the transmitted signal. That is, the coherence time of the channel is greater than the symbol period of the transmitted signal.

- **Fast Fading** is a kind of fading occurring with small movements of a mobile or obstacle that is caused by multipath phenomenon and environmental obstacles. These destructive or constructive interferences happen between the signal and its reflections and it varies over very short distances, in the order of a wavelength. In a fast fading channel, the impulse response changes rapidly within the symbol duration. That is, the coherence time of the channel is smaller than the symbol period of the transmitted signal.

[10] reviews various propagation models for both outdoor and indoor environments. Models, such as the free space propagation model, are used to predict signal strength when the transmitter and the receiver have a clear, unobstructed line-of-sight (LOS) condition. However, for most practical channels, the free space propagation model is too idealised, and inadequate to describe the channel and predict system performance. Commonly, the Rayleigh and the Ricean (Rice) propagation models are very popular in practical applications. For instance, the Rayleigh propagation model [11] allows for the situation when there is no LOS, and only multipath components exist. This model incorporates intensive variations in received signal power because multiple paths can either combine constructively or destructively. The amplitude, delay and phase shift of these components greatly depends on the environment. In the Rice propagation model not only models the situation when there is multipath effect, but also can take into account the effect of a LOS propagation path. For example, the Rice propagation model is used to model the environment within buildings where both LOS signal and multipath exist.

2.4 Wireless Location Technology

Many commercial applications such as navigation systems, health care systems and intelligent transportation systems adopt location information within their system designs. The positioning systems can be categorized by the measurement techniques to drive the

desired MS's location. A variety of wireless location techniques have been described in the literature. Traditionally, the major categories are based on the measurement of distance, angle, RSS-based, or any combination of the previous three categories. Figure 2.2 illustrates the basic classification of positioning techniques.

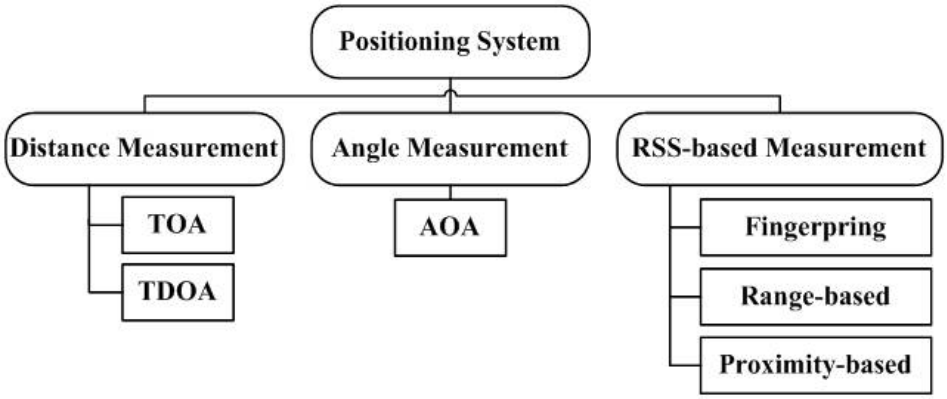


Figure 2.2: The category of positioning techniques

The main distance measurement approaches are known as TOA and TDOA. These two algorithms rely on the precision of timing between the signal transmitter and the receiver in order to use the propagation delay or time of flight (TOF) to calculate the distance between the transmitter and the receiver. Hence, a precise synchronization is needed in such a system. While the typical angle of measurement method (AOA) locates the MS by determining the angle of the transmitting signals. However, this approach requires the use of directional antennas and antenna arrays, which cause additional overheads. Although RSS-based localisation is typically less accurate than TOA-based positioning, it is still a very important technique since it can be implemented with little or no modification to existing systems. The use of RSS for location estimation is more economic and compatible for wireless networks because those related methods do not need any additional hardware, such as clocks or antenna arrays. In addition, RSS data can be readily collected indoors or outdoors in most wireless systems. The collected data can be used to obtain either range estimates, or connectivity information [12]. Fingerprinting, range-based and proximity-based localisation are the most common RSS-based localisation methods [13]. The description of each measurement technique is

as follows.

2.4.1 Distance Measurement Based on Time Delay

2.4.1.1 Time of Arrival (TOA)

The TOA technique determines a MS's position based on the measurement of the arrival time of a signal transmitted by a MS that is received at multiple BSs, as shown in Figure 2.3. It can be seen that if the signals emitted by the MS can reach a minimum of three BSs, the intersection point (as shown in Figure 2.3) is the estimated position on which the MS lies.

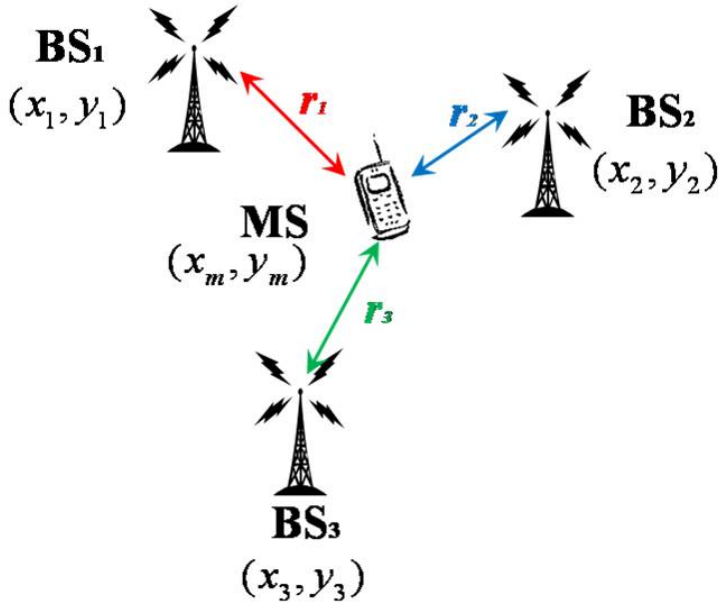


Figure 2.3: The time of arrival (TOA) method for MS location

In order to obtain the range between the MS and BSs, both transmitter and receiver should be equipped with clocks. If these clocks are synchronized, and if the start time of the signal transmitted by the MS is known, the TOA of the signal coming from the MS m to BS_i $(i = 1, 2, 3)$, $t_{m,i}$ can be measured (Here, m denotes the target MS). Then the distance r_i denotes the distance between the MS m and BS_i and it can be calculated as

$r_i = ct_{m,i}$, where c is the speed of light. According to the idea of tri-lateration [14], the distance r_i can also be expressed as finding the solution to the following equation:

$$\sqrt{(x_i - x_m)^2 + (y_i - y_m)^2} = r_i = ct_{m,i} \tag{2.1}$$

Here (x_i, y_i) is the coordinate of BS_i and (x_m, y_m) is the desired MS's position. Obviously, a small error in the clock at the receiver or the transmitter could result in large measurement errors in the distance prediction. Besides, multipath propagation and the case of non-line-of-sight (NLOS) conditions are another important source of error. Typical positioning errors caused by NLOS propagation in TOA-based techniques for GSM have been measured [15]. The reported average errors are in the range of 400 m to 700 m.

Since the measurement of TOA requires an accurate synchronization between transmitter and receiver clocks, it is too difficult to obtain relatively accurate timing information in mobile networks. So, many of the current wireless systems generally measure the round trip time (RTT) of a signal transmitted and then sent back in order to calculate the distances [16]. The RTT is the length of time it takes for a signal to be sent plus the length of time it takes for an acknowledgment of that signal to be received. This time delay therefore consists of the transmission times between the two points of a signal. In this case, the time of arrival can be simply obtained and the value is equal to half of the RTT. However, in the RTT method, the signal is sent both ways, which leads to additional overhead, and the processing delays at the mid-point of the round trip as the signal is sent back, which result in further timing uncertainty.

In the literature, a variety of TOA algorithms have been developed: the Taylor series expansion is utilized in [17] to acquire the location estimation of the MS from TOA approach by using iterative processes. It requires an initial guess of a MS's position, and then improves the guess at each step by determining the local linear least-sum squared error correction. However, due to an incorrect initial positioning guess, it may suffer

from the convergence problem, which may lead to incorrect results. Additionally the computational complexity of this approach makes it incompatible with some wireless applications. The Maximum Likelihood estimator is employed in the TOA system for more precise location estimation due to the NLOS situation. Based on the assumption of the Gaussian distributed TOA measurement error, the maximum likelihood location estimate is the globally optimal solution of a non-convex optimization problem. In [18], a two-step maximum-likelihood TOA-based algorithm is proposed. This approach aims to find the maximum-likelihood estimate of the MS location in a predefined restricted domain. A relative high accuracy positioning is provided when the non-line-of-sight propagation interference is not very heavy. In addition to the location estimation for a fixed MS, [19] [20] proposed a combination of Least-Square and Kalman filtering for location estimation and tracking.

Although the main principle of TOA and the techniques implemented are relatively easy to understand, it should be noted that the degree of precision of time synchronization between BSs and the MS strongly affects the accuracy of the MS location estimation.

2.4.1.2 Time Difference of Arrival (TDOA)

In the TDOA scheme, the location of a MS can be estimated regardless of the accuracy of the synchronization between BSs and a MS. It works by measuring the relative arrival time of the signals coming from at least three BSs at the MS at the same time, or by measuring the relative arrival time of signals emitted by the MS at the three BSs. Each time difference of arrival measurement can generate a hyperbolic curve, and the intersection of three hyperbolic curves is the location where the MS lies, as shown in Figure 2.4.

Figure 2.4 depicts the basic principle of TDOA. Assuming that the three BSs received a signal transmitted by the MS, and the estimated difference in propagation time from the MS to the BS_i and BS_k can be obtained by the use of a synchronized time reference

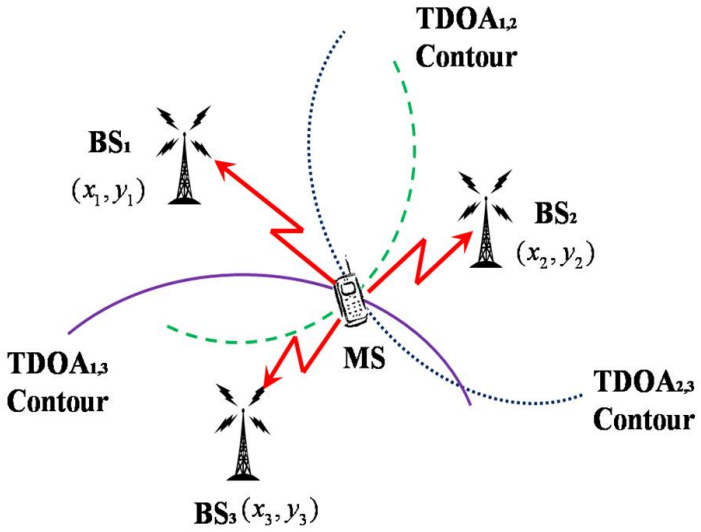


Figure 2.4: The time of difference of arrival (TDOA) method for MS location

among all of the BSs. Therefore, the difference in propagation distance to each BS can be calculated by using the simple distance equation which is given by [19],

$$D_{i,k} = ct_{i,k} = \sqrt{(x_i - x_m)^2 + (y_i - y_m)^2} - \sqrt{(x_k - x_m)^2 + (y_k - y_m)^2} \tag{2.2}$$

Where (x_i, y_i) and (x_k, y_k) represent the coordinates of BS_i and BS_k, respectively, $t_{i,k}$ is the difference between the TOAs of the MS signal at BS_i and BS_k, c is the speed of light, $D_{i,k}$ is difference distances from the MS to BS_i and BS_k, and (x_m, y_m) is the desired MS location coordinates. At least three BSs are required to perform the positioning location with TDOA. Due to the high nonlinearity of the set of equations above, many methods have been proposed to solve this problem. For instance, the least squares [21] approach is adopted to calculate the estimated position in TDOA measurement. Least squares works in a similar way to the maximum likelihood algorithm, but the computational efficiency of least squares makes it more popular than the maximum likelihood technique. Kalman Filtering is utilized in [22] and [23] to track the MS's trajectory based on TDOA in NLOS condition. The use of Kalman Filtering allows tracking the position and speed of the MS, yielding an accurate location prediction

algorithm. However, in the generally used Kalman Filtering in TDOA methods, the state transfer matrix is usually from the accelerated motion. That is to say, if the target MS is experiencing other kinds of motions, except for the accelerated motion, the state transfer matrix cannot reflect the real motion of the MS.

In a real environment, the utilization of TDOA in location estimation is more practical than the TOA system. It is independent of the signal emission time and does not require that all the system components be equipped with precisely synchronized clocks. The only requirement for the TDOA measurements is to ensure precisely synchronized clocks at the fixed location receivers (e.g. BSs). However, multi-path reflections, non-line-of-sight conditions, and other shadowing effects are the major factors that can lead to erroneous distance estimates.

2.4.1.3 Angle Measurement: Angle of Arrival (AOA)

The AOA technique is performed to determine a MS position by another triangulation technique [24] that utilizes triangle geometry in finding a location. In fact, it works in a similar way to TOA and TDOA methods, but instead of distances, the arriving angles of the signals coming from the MS to multiple BSs is measured. Generally, in order to obtain the desired location of the MS, the AOA approach utilizes antenna arrays at the BSs to measure the directions of arrival of the MS signals, then calculate and determine the potential position based on the intersection of directional lines of signals, as illustrated in Figure 2.5. As Figure 2.5 indicates, an estimate of the MS location can be obtained with only two BSs.

A wide variety of algorithms can be used to estimate AOA system. A typical approach uses beamforming techniques [25] [26] that focus on measuring the power spectral density between the antenna arrays to calculate the AOA of the MS signals. However, the accuracy of location estimation may be degraded by the limitation of the beam-width of the antenna array. Therefore, the accuracy of AOA location estimation might be

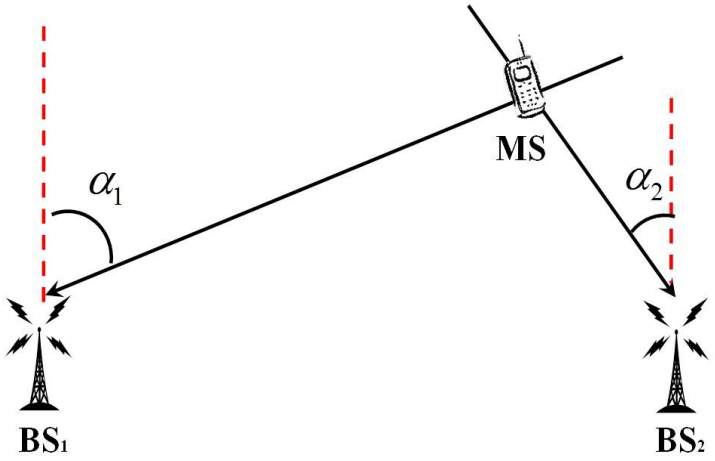


Figure 2.5: The angle of arrival (AOA) method for MS location

strongly affected by the multipath environment, and also it performs better for an open environment than an indoor environment. In a dense multipath environment, multiple delayed copies of a signal may arrive from different directions, which could make most input signals highly correlated. These multiple signal copies might be caused by signal reflections, scattering from the buildings, trees, etc. Maximum Likelihood (ML) direction finding algorithm [27] is used to resolve closely spaced correlated signals. The algorithms in [27] require a multidimensional search, but efficient techniques for performing such a search have been developed. Another approach is to use spatial smoothing followed by some high resolution algorithm such as the MUtiple Signal Classification (MUSIC) [28] algorithm. This approach is less computationally intensive but it does not perform as well as an ML technique. Besides the algorithm selected, factors which affect direction finding accuracy include signal-to-noise ratio, integration time, number of antennas, hardware nonidealities, and array calibration error. What is more, with the increasing demand for higher bit rates in wireless communications, Multiple Input Multiple Output (MIMO) systems are being considered as a promising way to improve the capacity of the radio channel with minimal frequency resources [29]. These systems use multiple antenna elements at the transmitter and receiver to improve the capacity over Single Input Single Output (SISO) systems when operated in multipath environments. Especially, in the non-line-of-sight areas, the channel correlation and the channel capacity are influenced

by the distribution of the incoming signal such as power azimuth spectrum, AOA and direction of arrival. So, estimating direction of arrival is essential in MIMO performance measurement.

Unlike time-based methods, the AOA location approach does not rely on precision timing between BSs and the MS, and also it does not require a high accuracy clock in the communication system. Moreover, at least two BSs are needed for the AOA location process. The minimum number of BSs for the AOA system is less than the time-based techniques (TOA and TDOA), which require three BSs. However, the AOA techniques have some drawbacks. Firstly, these require relatively complex and expensive hardware to measure the direction of arrival of the MS signals. Furthermore, the AOA accuracy strictly depends on the network topology and propagation environment, the prediction accuracy can be seriously degraded due to noise, fading, and interference in physical wireless environment. The accuracy of the AOA method diminishes with increasing distance between the MS and BS due to the scatter environment and fundamental limitations of the devices used to measure the arrival angles. [19] gives an example and consider a scenario in which a measured AOA is in error by 3° at a certain BS. A MS is located 200 meters away from the BS will be 10 meters away from the line of position, while a MS located 1000 meters away will be located 52 meters away from the line of position. This leads to larger error for the further MS since the location estimate is determined by the intersection of the line of position.

2.4.2 RSS-based Localisation

For many practical location applications, the goal of a location system designer is to minimize the system requirements despite reasonable degradation in location accuracy. Therefore, the RSS-based approach is an attractive candidate for location estimation in wireless networks. Despite being less accurate than TOA-based positioning, RSS-based localisation is seen as simple, economic for wireless networks as it does not require additional hardware and can be applied without the aid of a network operator or third-

part of data.

Two major aspects should be taken into consideration in relation to RSS systems: the accuracy of the location estimation and the range of the radio coverage area. For RSS-based location systems, the primary source of error is multipath fading and shadowing [30]. A number of empirical models have been set up [10], but they depend on both the signal properties and the environment where measurements are performed. Moreover, many studies have developed different algorithms utilizing RSS measurement for position prediction in the past few years, such as clustering methods [31], particle filtering [32] and probability-based algorithms [33]. With the author's understanding, the RSS-based localisation methods can be broadly classified into three categories: fingerprint localisation, range-based localisation and proximity-based localisation. Among them, the fingerprinting technique is the most widely used technique for positioning.

2.4.2.1 Fingerprint localisation

Fingerprint localisation could overcome the limitations of traditional triangulation approaches and perform well for NLOS circumstances, especially in a complex environment. For a typical fingerprint-based localisation system, it only needs to collect the measurements of RSS or other useful parameters at some known locations to form a location fingerprints database (a.k.a. radio map) during the training stage. When a new online RSS tuple is observed, its location is estimated using various algorithms based on the location fingerprints database. Chapter 3 will give a detailed description of fingerprinting localisation techniques and systems.

2.4.2.2 Range-based Localisation

Range-based localisation algorithms assume that the signal strength decays over distance following a distribution that is known a priori. This distribution is used for converting one or several signal strength measurements into distance estimates. Generally, these

distributions include several parameters that try to account for the influence of the environment which are calibrated in the calibration phase [34] [35] [36].

2.4.2.3 Proximity-based Localisation

Proximity based localisation algorithms assume that the signal strength decays inversely proportionally with the distance [37] [38] [39]. The main difference between proximity-based algorithms and range-based algorithms is that proximity-based localisation only uses the order of RSS measurements instead of converting signal strength to distance estimates. The advantage of proximity based localisation algorithms is that they do not require a calibration phase.

2.4.3 Performance Comparison of Location Techniques

Table 2.1 compares the four classical methods previously discussed in terms of the number of BSs required for localisation, the need for LOS, the application environment

Table 2.1: Comparison of the Basic Measurement Methods

Methods	**No. of BSs**	**LOS / NLOS**	**Environment**	**Accuracy**	**Extra Contribution**
TOA	≥ 3	LOS	Outdoor	High	Time synchronization across transmitter and all receivers is needed.
TDOA	≥ 3	LOS	Outdoor	High	Time synchronization at all receivers is required.
AOA	≥ 2	LOS	Outdoor	Low	Smart antennas are needed.
RSS	≥ 3	Both	Both	High to Medium	Simple and inexpensive.

(whether the technique is applied to indoor or outdoor), localisation accuracy and extra requirements for the positioning. For time-based techniques, both TOA and TDOA measurements strongly rely on the precision of the timing between the signal transmitter and receiver. Thus, a high accuracy clock plays an important role in wireless location estimation systems. To some extent, TDOA is preferred to TOA in many implementations. The TDOA estimation only needs the clock synchronization between the BSs in the system. AOA has the advantage over TOA and TDOA as neither BSs nor the MS is needed be synchronized. However, specialised and complex antenna hardware is required and the location accuracy forcefully decreases in larger cells. In real scenarios, hybrid techniques use more than one type of location estimation method to improve estimation accuracy. Typical examples are TOA/AOA [40], TDOA/AOA [41] [42] and so on. Despite the relatively low accuracy for the positioning, the advantages of RSS over the other three approaches rely on the fact a) it does not need to have additional hardware or a particular network operator support to measure parameters; c) the RSS data is readily collected and monitored and performs well in non-line-of sight circumstances. Hence, the RSS-based approaches have been widely investigated principally in the context of indoor location estimation. In addition to the geometric-based location estimation for a MS, some machine learning methods and probabilistic approximation algorithms have been proposed and employed in the classification of MSs or the location tracking of a MS.

2.5 Location Determination Systems

[43] presents a comprehensive survey of classical positioning systems. Therefore, this section will take a small subset of these systems as examples and briefly introduce their major characteristics.

2.5.1 GPS

The Global Positioning System (GPS) is probability the most widely known positioning system. The GPS satellite constellation consists of 24 satellites, each with an orbit of 12 hours [44], so that almost all users on the earth can see at least four satellites simultaneously, anywhere on the globe and at any time. It employs signal timing based upon the same principle as TOA to measure the distance from the satellites to the user receiver, and then determines the target location of the user.

In fact, many engineering or commercial companies utilize GPS technology for their own applications, especially in vehicle-based systems or mobile phones. For instance, GPS navigation installed in cars can be used to guide drivers through a detailed route to their destinations. However, it is seen that GPS is unlikely a best option to solve positioning in cellular networks. The reasons include: a) In order to obtain accurate TOA measurements, GPS requires precise time synchronization between satellites and receivers and so the receiver clocks bias has to be accounted for. b) Multipath reflections make the satellite signals become weak when they arrive in cluttered or indoor environments, consequently GPS provides an inaccurate location. Hence, GPS needs LOS, and it is unusable in dense environments such as urban environments with many tall buildings.

2.5.2 AGPS

When the GPS system is first turned on, it needs a long time (e.g. from 30 seconds to a couple of minutes) to acquire satellite signals, to navigate data, and to localize. This problem is called time to first fix (TTFF) or "cold start" [8]. The time duration depends on the location of the receivers and the surrounding interference and horizon information. Therefore, the Assisted Global Positioning System (AGPS) has been developed in order to combat this shortcoming of GPS.

The main system components of AGPS are a wireless handset with partial GPS receiver, an AGPS server with a reference GPS receiver that can simultaneously monitor and track the same satellites as the handset, and a wireless network infrastructure consisting of BSs and a mobile switching centre [8].

The AGPS server obtains the handset position from the mobile switching centre and can locate the cell of the handset and even the sector of the handset within a set if directional antennas are used at the base stations [45]. Because the AGPS server monitors and tracks the GPS satellites, it can predict the satellites that are sending the signals to the handset at any given point of time. Thus, the AGPS server can communicate the satellite information to the handset. This enables the handset to acquire GPS signals quickly when it is first turned on, reducing the TTFF from minutes to less than a second. Once the satellite signals are acquired by the handset, it calculates the distances to satellites without clock synchronization. These satellite distances are sent back to the AGPS server for further computation. Therefore, the AGPS server also shares the computational load of the handset, reducing the handset battery power consumption. Despite the fact that AGPS can improve the performance of a conventional GPS, the indoor positioning problem is still not satisfactorily resolved, and also the extra signalling needed from the GPS reference stations increases the impact in a mobile wireless system.

2.5.3 RFID

The RFID, Radio Frequency Identification, is a wireless system that identifies tags attached to the object of interest. An RFID system consists of a reader and RFID tags. RFID systems are divided into two categories, according to whether they use passive or active tags [46]. For the passive tags, they are suitable for short-range application because they do not contain a power source. The passive RFID tags are equipped with an antenna that is excited by output signals at specific frequencies, and these tags are activated by the power of the received signal. While for the active RIFD system, it is a full transceiver system including processors, antennas, and batteries. Thus, an active

tag contains both a radio transponder and a power source for the transponder. An RFID reader constantly sends Radio Frequency (RF) electromagnetic waves, which are received by the RFID tag in its vicinity. The RFID tag modulates the wave adding its identification information and sends it back to the reader. The reader converts the modulated signal into digital form to determine the tag identity. Active tags are ideally suitable for the identification of high volume products moving through a processing unit [8].

The RFID can be used to localize the position of a target object. An active RFID tag can be attached to the object, which transmits a signal to the RFID reader. The concept of trilateration is used along with the received signal strength indication technique to localize the position of the tag. Because the objects to be positioned using RFID are usually in an enclosed environment, there are multipath effects, which decrease the accuracy of the system. In order to increase the accuracy of RFID-based positioning system, the system utilizes additional readers and reference tags. However, these additional readers increase the cost of the system. In order to keep the costs down, [47] proposed an innovative approach, LANDMARC, that employs the idea of installing extra fixed reference tags. This approach is called location identification based on dynamic active RFID calibration.

2.6 Radio Coverage Prediction

With the rapid growth and demand for high quality and high capacity networks, providing sufficient radio coverage with minimum cost has become extremely important. For this purpose, it would be useful to have the ability to accurately predict radio coverage area and traffic demand behaviour for various scenarios, which can reflect topographical features and dynamic network reconfigurations. Inaccurate estimation of radio coverage has severe impact on the network performance. Over estimating network coverage will cause “coverage holes” where the areas with signal strength are weaker than the

minimum required threshold. Under estimating network coverage will create coverage overlaps, which result in interference. Hence, accurate prediction of radio coverage is essential for network planning.

As a result of heterogeneous traffic demand, hard-to-predict users' movements and complex propagation models, how to achieve flexible radio coverage in a realistic environment is a tough challenge for network providers.

Cooperative control is an effective way to handle heterogeneous traffic situations and to improve the whole network performance significantly. The main idea of the RF cooperative control is to use the RF domain optimization to increase utilization of the limited frequency spectrum at reasonable costs. In other words, cooperative control could cut the cost and time spent on network deployment by simultaneously optimizing the RF domain according to perceived traffic demand distribution and propagation environment. Thus, cooperative control can exploit the potential flexibility of wireless networks to respond to traffic load demands in the RF domain of the physical layer and optimize the network performance.

Previous research [6] [7] [48] on cooperative control proposed to create flexible radio coverage according to different networks and antenna models. Their results have shown significant improvements in call blocking, call dropping and system capacity. The cooperative control method adjusts the physical layer, such as the radiation pattern, titling angle or transmit power based on the traffic demand in collaboration with other cells. The bubble oscillation algorithm is introduced [48]. The main concept of this approach is to use an analogy with air bubbles: the local coverage scheme is treated as an air bubble, the local traffic load is treated as the air within the bubble and the un-served traffic is treated as a vacuum between adjacent bubbles. The bubbles oscillate to obtain optimum radio coverage across the network. In [7], a statistical model is proposed for radio coverage prediction based on the received signal power feedback of the MSs to improve the accuracy of radio coverage and avoid network holes in a realistic environment. With a distributed approach [6] and the additional capabilities of evolved Node B(eNB), cooper-

ative control can result in self-optimization for LTE networks. In order to reason about the best cooperative coverage in novel situations, it is crucial to propose and develop an approach, which can build different models of the expected RF environment. These approaches depend on a way to predict coverage in different configurations, and so the work in this thesis is central to the eventual exploitation of such theoretical work. During the last years, a couple of new propagation models for indoor and outdoor prediction have been proposed in the literature, e.g. [49] [50] [51]. In [49], a variety of experimentally or theoretically based models have been introduced to predict radio propagation in land mobile systems. [50] presents a run-time efficient three-dimensional propagation model for the complete prediction of outdoor and outdoor-to-indoor coverage of small macro cells in urban areas based on high-resolution building data.

2.7 Summary

To better analyse the research work, related background about network radio characteristics and localisation methods were investigated in this chapter. Firstly, the basic idea of radio network planning was described followed by the overview of radio propagation model. The technical aspects of wireless location technology and existing location estimation systems were then discussed. Finally, the challenges and previous studies on radio coverage prediction were discussed. The next chapter will address the fundamental aspects of fingerprint location estimation based on RSS measurements.

Chapter 3

Received Signal Strength-based Fingerprinting Localisation

3.1 Introduction

The use of RSS is ubiquitous in wireless systems. While RSS-based localisation is typically less accurate than TOA-based positioning, it is still a very important technique since it can be implemented with little or no modification to existing systems. Thus, this chapter reviews the literature on location fingerprinting by using signal strength. In section 3.2, the basic concepts of location fingerprinting are considered. Previous work related to transmitter selection is surveyed in section 3.3. Section 3.4 discusses different means to create the radio map, followed by an analysis of the three main characteristics of location fingerprinting: the environment feature, the partitioning models and the location fingerprinting techniques in section 3.5, section 3.6 and section 3.7 respectively. The conclusion of this chapter is presented in section 3.8.

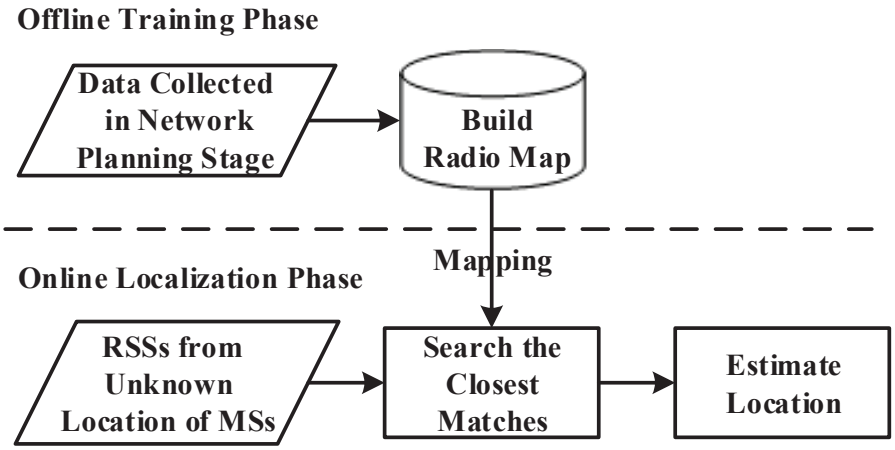


Figure 3.1: The structure of fingerprinting method

3.2 Location Fingerprinting

Location fingerprinting is the positioning technology that utilizes the relationship between a particular MS location and its corresponding RSS tuple value. Unlike other localisation techniques, e.g. TOA, TDOA and AOA, that require the calculation of the distance between the target MS and the BSs, and the triangulation of the MS's location, location fingerprinting techniques usually only need to measure the RSS at certain locations to build up the "fingerprint" database and decide the location of a MS by comparing the obtained RSS to the database. Generally, the procedure of location fingerprinting can be divided into two phases: *offline training phase* and *online localisation phase*. Figure 3.1 illustrates the structure of location fingerprinting. The arrows in this figure show direction of flow from one step to another.

In the training phase, the relative fingerprints are required to be collected and stored in a database. Fingerprints can be gathered by performing a site survey of the RSS from multiple transmitters, e.g. BSs or RSs or APs, in the target environment. The RSS is measured at every predefined location to create a database of RSS patterns of this area. The generated database of RSS pattern is called a *Radio Map*, which consists of many Location-RSS vector pairs. Every RSS vector is called the location fingerprint of its corresponding location.

In the online estimation phase, the new RSS observation originated from the unknown positions are used to compare with all the fingerprints stored in the database during the data collection period, to estimate the location based on different kinds of algorithms. The most common and simple method is to calculate the *Euclidean distance* between the new RSS vector and all the fingerprints to find the most similar RSS pattern, and then the location of this best match pattern is returned as the estimated location for the new RSS vector.

As discussed, a radio map created during the training phase covers the area of interest and contains the RSS information about the function of the location, and then it is used as a reference to determine the estimated position further. This makes location fingerprinting methods reliable in relatively complex environments. Therefore, in order to ensure a high precision of location estimation at run time, it is very significant to design an appropriate radio map in the location fingerprinting process.

However, several important issues need to be considered: firstly, to achieve a good estimation of MS location, the more training data collected during the offline phase, the better; the more measurements obtained at each training data, the better, which means it is a significant, time-consuming task during the data collection phase, especially performing a huge data collection for a wide area network. Furthermore, it is a problem of choice as to how many transmitters are required for the system and how to pre-process the raw data before any further operation. What is more, how to keep the radio map up-to-date in the constantly changing real environments, especially in urban environment also needs to be taken into account.

3.3 Transmitter Selection

It makes intuitive sense that the more transmitters, e.g. BSs or RSs or APs, used in the location estimation system then the higher the positioning accuracy obtained. This is because the signal strength from any transmitter can provide some information for

location estimation. However, the use of more transmitters increases the dimensionality and computational complexity. Hence, how to select the BSs to avoid unnecessary calculations, minimize noise levels and achieve high position accuracy is one of the challenges in this work.

[31] [52] show how their smart AP selection methods can achieve good localisation results as compared with using all the available APs in an indoor environment. In [31], the MaxMean approach is proposed to choose the K most important APs. These are defined to be those K APs for which the average RSS is the highest. This mechanism unavoidably throws out the information of detectable but unselected APs, and also requires that at least one AP can communicate with every point in the grid. This makes the approach only suitable for small areas. [52] introduced the InfoGain algorithm for AP selection, which divides the indoor environment into n grid elements. Let G_j denote the $j-$th grid element in the test-bed. Suppose m is the number of APs detectable. The signal strengths from the APs are collected in every grid G_j. The average value of signal strength in G_j from AP_i $(i = 1, ..., m)$ is defined as the value of the i-th feature of G_j. The main idea of InfoGain is to select the top K APs in terms of the "worth" of each AP feature in every grid element. The worth of each AP_i feature is calculated as the reduction in entropy by including the feature, which is given by $InfoGain(AP_i) = H(G) - H(G|AP_i)$. Here $H(G) = -\sum_{j=1}^{n} Pr(G_j) \log Pr(G_j)$ is the entropy of the grid when AP_i's RSS value is not known, $Pr(G_j)$ is the prior probability of grid G_j and is treated as uniformly distributed, i.e. a user can be equally likely in any grid [52]. $H(G|AP_i) = \sum_v \sum_{j=1}^{n} Pr(G_j, AP_i = v) \log Pr(G_j|AP_i = v)$ computes the conditional entropy given AP_i's value. v is one possible value of signal strength from AP_i. The summation is taken over all possible values of AP_i. So they only need to focus on the value of $H(G|AP_i)$ which is the conditional entropy of grid G given AP_i's RSS value. Although positive results have been demonstrated for a relatively small indoor environment, for a larger area, such as an outdoor environment, it is difficult to determine the appropriate number of grid elements for the target area and the size of each grid

element.

3.4 Radio Map

In the location fingerprinting system, the positioning accuracy is affected by a mapping between the signal space and physical location space. The design of a fine-grained radio mapping is an important part of the location fingerprinting process. The radio map describes the signal distribution for all wireless networks receivable in the target area, which can reflect how strong the reception of each network is for each of several discrete locations (generally called reference points). These reference points are spread over the covered area and can be described by their geographical coordinates. The signal distribution patterns for every reference point captured during the training phase are referred to as fingerprints. Generally, each fingerprint comprises of several data records, which represents the received signal strength of one specific network at this reference point. Additionally, the radio map is the union of all recorded fingerprints. During the online phase, an algorithm evaluates the fingerprints stored in the radio map and determines a certain distance between the stored and new captured signal patterns.

Different ways to generate a radio map have been proposed in previous studies by other researchers. The simplest method is to store only the mean value of RSS measurements at each training sample [53], regardless of the variation of the RSS. [54] proposed a method based on interpolation to create a radio map. In [55], the author surveys the properties of RSS in relation to the performance of a positioning system and provides guidelines on the design of indoor positioning systems. Furthermore, [56] investigates questions related to the design of IEEE 802.11b wireless LAN networks (WLAN) radio maps for localisation based on fingerprinting, They analyse and compare the effects of design factors, e.g. the size of the radio map and auxiliary supporting information in the radio map. In this thesis, RSS is the main parameter for location estimation.

3.5 Indoor and Outdoor Environments

Existing localisation systems can be broadly classified into *indoor* and *outdoor* localisation systems according to their environmental dependency. Accordingly, this thesis analyses the difference between indoor scenario and outdoor scenario in three aspects:

Firstly, scale is an important parameter that impacts how fingerprints can be collected and exploited effectively. Positioning systems need to scale on two axes: the size of the land area and the population density of that area. For indoor localisation systems like in-building navigation and digital homes, the typical environmental situations for these applications are at home, in an office room or block or in a shopping mall, which typically have small or medium sized areas with low-density of mobile users. Though in the case of a shopping mall there may sometimes be a high density of users. In contrast, outdoor localisation systems such as GPS, locate and track mobile users in the street, in the park, or along the river and often cover larger-scale and high-density deployment.

Secondly, due to the smaller size of area and the nature of the target applications, the accuracy requirement for indoor positioning systems is often relatively higher than for outdoor positioning systems. However, both indoor and outdoor radio propagation scenarios suffer from multipath fading, shadowing effects and interference to different degrees. The outdoor environment can be divided into urban, sub-urban and rural areas based on the type of terrain. Each of them has specific features that influence the positioning system performance. For instance, the strong effect of the canyon phenomenon in urban areas leads to high variability of RSS between geographically close points [57]. This is the main contributor to accuracy degradation in triangulation location methods. Generally, the localisation accuracy depends on how much the path loss model reflects the realistic propagation environment.

Moreover, the mobility of the user greatly depends on the environment. In general, low speeds and well-defined mobility paths are characteristic of indoor environments, whereas variable speeds and flexible mobility paths that depend on the particular envi-

ronment are features of outdoor environments. For example, there is a increase in the predictability of user mobility when a user is walking or driving in a dense city environment with fixed streets and pathways relative to when the use could wander anywhere in a rural village, field or forest.

3.6 Grid-based versus Cluster-based Localisation

To improve the accuracy of location estimation and reduce the computational load, most previous localisation schemes have been built based on models that partition the environment. *Grid-based partitioning* and *cluster-based partitioning* are the two most popular in localisation systems.

The positioning applications based on grid partitioning [58] [59] generally divide the simulation environment into a uniform regular grid and attempt to map a MS location to a point on a grid element. The spacing of the grid influences the accuracy of the position estimate [54]. A key issue is that a uniform grid does not reflect the topography. Choosing larger grid spacing reduces the accuracy of positioning dramatically. On the other hand, choosing smaller grid spacing increases the accuracy but leads to a more laborious site-survey. Some location-aware applications [58] (mainly indoor ones), do not use a regular grid but use a topographical model of the environment, where the environment, e.g. an office building is divided into cells where each cell corresponds to a specific office room or hallway segment, in the office building. Although these grid-based localisation techniques partly improve the localisation accuracy, the selection of grid size is only loosely related to the radio propagation conditions in different areas.

As a response to this, many cluster-based location estimation methods have been proposed in recent literature. Again most of them have concentrated on the indoor environment and WLAN [31] [52] [60]. Those results have shown that a cluster-based structure is a good prediction tool for locating and tracking the MSs. The clustering scheme is used to partition the environment into geographic regions that are homoge-

nously covered by the radio signal. The aim is to better model a complex environment. In [31] [60], the authors present the Joint Clustering technique to cluster the locations in the radio map based on covering APs during the offline phase in order to decrease computational complexity, and then apply a Maximum Likelihood (ML) estimator to determine the most probable location within the cluster during the online phase. [52] proposed a new algorithm known as CaDet for power-efficient location estimation by selecting the number of APs used for location estimation in an indoor wireless environment. The simulation environment is modelled as a space of 99 locations, each representing a 1.5-metre grid cell. In the offline mode, it uses K-means clustering [61] to generate clusters based on the similarity of signal strengths from APs. In the online mode, the decision tree over the grids in each cluster is built to detect a target's location with high accuracy. There are two points worth noting. Firstly, previous clustering localisation research has not paid attention to the cluster stability and scalability issues associated with handling a large amount of data without losing important correlation information. Secondly, other approaches to clustering are dominated by the path loss effect and the clusters are determined to a large degree by this, rather than by the correlations between the RSS values created by topographical effects. The method used in this thesis makes an approximate adjustment for the distance effect and then works with the residuals[2], and also has the benefit that the clusters are invariant to the power at the BSs and RSs.

3.7 Estimation Techniques

As discussed previously, in the online location estimation phase, new RSS measurements are compared to the radio map to calculate the predicted location. Accordingly, the fingerprint-based approaches are generally classified into two categories: *deterministic techniques* and *probabilistic techniques* depending on how a location estimate is generated.

[2]Residual is represented the difference between an observed value and its estimated value from a regression model.

3.7.1 Deterministic Estimation

Deterministic techniques use deterministic inference algorithms to estimate a MS location. This essentially involves calculations of the similarity between new RSS observations and the training RSS samples that are associated with known location information.

The RADAR system [53] [62], a RF based system for locating and tracking users inside buildings, represents the first 802.11 fingerprinting structure for localisation developed and was by Microsoft Research. The system uses K-Nearest-Neighbour (KNN) algorithm [63] to estimate the desired user's location as the average of the coordinates of the K training locations whose fingerprint tuples have the shortest Euclidean distances to the online RSS tuple. [47] and [64] improve the location determination accuracy by using a weighted average of the coordinates of the K nearest training samples. The weight values are taken as the inverse of the Euclidian distance between the target RSS measurement and the RSS measurements of the K training samples. This method is referred to as Weighted K-Nearest Neighbours (WKNN). The experimental results in [64] indicate that the KNN and the WKNN can provide a relatively higher accuracy than the Nearest Neighbour (NN) method, particularly when $K = 3$ and $K = 4$. However, when a high density radio map is available, i.e. there is a lot of training data, the simple NN method can perform as well as other more complicated methods [65]. Moreover, some variants of the KNN method, e.g. the Database Correlation Method [66] [67], have been explored to predict the location.

3.7.1.1 Distance Measurement in Signal Space

Suppose that the radio map constructed during the training phase consists of a set of n location fingerprints denoted by $\{r_1, r_2, ..., r_n\}$ and each fingerprint has a one-to-one mapping to a set of positions $\{l_1, l_2, ..., l_n\}$. A sample of an RSS fingerprint measured during an on-line phase is denoted as r_m. Assuming that this target environment only considers the RSS from q transmitters, e.g. BSs or APs or RSs as a location fingerprint,

the sample of RSS vector is $r_m = (r_{m,1}, r_{m,2}, ..., r_{m,q})$ and each RSS fingerprint r_i in the radio map can be denoted as $r_i = (r_{i,1}, r_{i,2}, ..., r_{i,q})$.

There are different ways to find the best match between the RSS observations and radio map. The common choice for the comparison measure is to use the *Euclidean distance*. Let $S(\cdot)$ function be the a distance measurement in signal space. According to the Euclidean distance [53] [68] [69], the similarity between the new observation r_m and RSS fingerprint r_i from the q transmitters can be expressed as

$$S(r_m, r_i) = \sqrt{\sum_{k=1}^{q} (r_{m,k} - r_{i,k})^2} \tag{3.1}$$

The Minkowski Distance [70] is a generalization of the Euclidean Distance, which can be given by

$$S(r_m, r_i) = (\sum_{k=1}^{q} (r_{m,k} - r_{i,k})^p)^{\frac{1}{p}} \tag{3.2}$$

where p is the norm parameter. Starting from 1 and by varying parameter p, different norms can be obtained. For example, $p = 1$ corresponds to Manhattan distance and $p = 2$ implies Euclidean distance in (3.1). [68] modified (3.2) and added the weight factor $w_{i,k}$ in (3.2), that is, $S(r_m, r_i) = (\sum_{k=1}^{q} \frac{1}{w_{i,k}} (r_{m,k} - r_{i,k})^p)^{1/p}$,where $p = 2$. In [68], the weight $w_{i,k}$ is assigned to the number of RSS samples or the standard deviation of RSS fingerprint. The weight is considered as bias parameter that can demote or promote an important RSS component in the fingerprints [65].

In addition, the *Mahalanobis distance* [71] can also be used as a distance measure. In this thesis, Mahalanobis distance is used rather than Euclidean distance, because it can automatically account for the scaling of the coordinate axes, correct for correlation between different features, and enable both non-linear and linear decision boundaries [71]. Direct comparisons of the Euclidean distance and the Mahalanobis distance on real data sets are given in chapter 5 section 5.5. In fact, the Euclidean distance is a special case of the Mahalanobis distance when all the RSS signal components in the

location fingerprint are uncorrelated and their variances are the same in all directions. However, the disadvantage of the Mahalanobis distance is that the covariance matrix for the location fingerprint must be determined.

Given a location fingerprint vector r_i, a sample vector r_m, and a covariance matrix of location fingerprint Σ, the Mahalanobis distance $S(r_m, r_i)$ can be calculated as

$$S(r_m, r_i) = -\sqrt{(r_m - r_i)^T \Sigma^{-1} (r_m - r_i)} \tag{3.3}$$

3.7.1.2 K-nearest Neighbour

The K-Nearest Neighbour (KNN) [63] method has been widely used as a benchmark in localisation research. Previous research work have shown that the KNN method can provide good accuracy if enough training data are collected during the training phase.

In the KNN approach, the estimated location $\hat{l}$ is calculated as the average value of the K training data's locations with the smallest RSS distance between the new observation r_m and RSS fingerprints r_i in the q-dimensional RSS space (e.g. q is the number of BSs) stored in the radio map, which can be expressed as

$$\hat{l} = \frac{1}{K} \sum_{i=1}^{K} l_i \tag{3.4}$$

Where the set of $\{l_1, l_2, ..., l_K\}$ denotes the ordering of reference locations with respect to increasing RSS distance between the respective r_i and the observed RSS measurement r_m.

3.7.2 Probabilistic Estimation

Probabilistic techniques are often used as a tool for modelling uncertainty and errors in RSS measurements in wireless networks [72]. These methods use the training RSS

samples to construct the probability distribution of RSS over various locations as the content of a radio map, and then utilize probabilistic inference to compute the likelihood or posterior probabilities over locations during positioning. The MS localisation can be derived from the likelihood and posterior density function by using the ML estimator [31] [60], the Minimum Mean Square Error estimator [72], the Maximum A Posterior estimator, or variants of those estimators (e.g. Probability-based Maximum Likelihood [33]).

The estimation of unknown probability density function (pdf) from the training RSS measurements constructed in the offline phase is a problem of fundamental importance to positioning accuracy. Density estimation methods can be roughly categorized into parametric and nonparametric. In parametric density estimation, the pdf is approximated as a certain particular known distribution function, e.g. the Gaussian distribution [58]. However, due to the complex propagation environment, the distribution of RSS can be asymmetric and multimodal [65] [73]. Accordingly, it is difficult to build up a parametric probability distribution with a known function to RSS from the real environment. On the other hand, a nonparametric approach, such as the construction of the RSS histogram from sample data or a kernel density estimator (KDE), can estimate the pdf without assumption regarding the specific form for the density. Although the histogram estimation [72] [74] is the most widely used to estimate the pdf, the constructed density estimate is highly dependent on the choice of starting position and bin width [3] . Additionally, the histogram is discontinuous because of the use of discrete bins and sensitive to disturbances due to the sample size for areas with low sample frequencies. All these problems make it unsuitable for most practical applications, especially in higher dimensional spaces. The KDE [72] [75] [76] is considered as a more general density estimation technique than a raw histogram approximation, and can smooth the discrete histogram to a continuous function and the smoothing accommodates incomplete RSS data. The authors in [72] point out that the accuracy of the KDE is highly dependent on the number

[3]To construct a histogram from a continuous variable, it needs to split the data into intervals, called **bins**, and the size of bin is **bin width**.

of training samples and the size of bandwidth[4] (more detail will be given in chapter 6). Furthermore, some kernel-based methods have also been explored for positioning, like the Support Vector Machines [77] and Canonical Correlation Analysis [78] [79]. In addition, Bayesian inference provides a general framework for positioning and tracking problems. For example, the Nibble system [80] applies Bayesian networks with RSS measurement to infer the location of a target sensor in a Wi-Fi network environment. [81] provides a detailed survey of the application of Bayesian Filtering for location estimation.

3.7.3 Comparison of Estimation Techniques

Some classical localisation schemes in the literature have been classified with respect to different categories as shown in Table 3.1. The procedures used in deterministic techniques are relatively simple, but generally need to collect a large number of training data during the offline phase in order to achieve high location estimation accuracy. Although probabilistic techniques are reported to provide higher positioning accuracy than deterministic techniques, and this has been proven in [72] based on their test-beds, their higher computational complexity makes probabilistic techniques difficult when the vector of observations is of high dimension.

In this thesis, different approaches have been proposed. For outside environment, two location approaches are developed to estimate a user's location: the Intersection after Principal Component Analysis (PCA-Intersection) method and the Kernel density estimator after PCA (PCA-KDE) method, both of which are augmented with the use of PCA. The first one belongs to the deterministic category, and the second one is within the probabilistic category. For inside environment, the deterministic location approach called the Weighted K-Nearest Neighbour after PCA (PCA-WKNN) is utilized to predict the room number. For all of the three approaches, the PCA method is used for feature selection to reduce the required data, as well as to retain important correlations in high dimensional data in the reduced dimension. The benefit of PCA is that it reduces the

[4] **Bandwidth** is a smoothing parameter that controls the smoothness of the density estimation.

Table 3.1: Classification of Some Localisation Schemes

Localisation Scheme	**Area of Deployment**	**Partition Model** Grid/Cluster/Global	**Estimation Techniques** Deterministic/Probabilistic
RADAR [53] [62]	WLAN Indoor	Global	Deterministic
Ref. [72]	WLAN Indoor	Global	Probabilistic
Ref. [64]	Indoor	Global	Deterministic
Nibble [80]	WLAN Indoor	Global	Probabilistic
Ref. [58]	WLAN Indoor	Grid	Probabilistic
CellSense [59]	GSM Indoor	Grid	Probabilistic
Horus [60]	WLAN Indoor	Cluster	Probabilistic
CaDet [52]	Indoor	Cluster	Probabilistic
PCA-Intersection	GSM Outdoor	Cluster	Deterministic
PCA-KDE	GSM Outdoor	Cluster	Probabilistic
PCA-WKNN	Indoor	Cluster	Deterministic

computational requirements of location determination and can perform simple lookups with fewer samples.

3.8 Summary

In this chapter, the general aspects of location fingerprinting systems were described, followed by a survey about its four main characteristics that may influence the precision of localisation accuracy: transmitter selection approach, indoor or outdoor environment, grid model or cluster model and deterministic or probabilistic method applied to positioning system. The chapter concluded with a short evaluation of location fingerprinting approach. The following chapters, Chapter 4 to Chapter 8, concentrate on the proposed outdoor localisation techniques and coverage prediction. These chapters present the proposed run-time positioning measurement system in outdoor environment and offer different approaches to estimate users' locations in a static/dynamic environment as well as prediction the coverage reliability.

Chapter 4

Partitioning the Wireless Environment

4.1 Introduction

This chapter describes the approach to partitioning the environment into different disjoint regions where each region in RSS space maps the locations in a real environment that have similar RSS. This is achieved by creating clusters in the space of RSS. The clusters do not necessarily correspond to physically contiguous regions. The MS belonging to a cluster can be dispersed over the geographical area and even be interspersed by MSs that belong to different clusters. So a MS at any particular geographical location may belong to more than one cluster. There is a trade-off between the number of clusters and the accuracy of the location estimation. The fewer the number of clusters the more accurately the cluster can be identified, but the less useful the identification of the cluster is. This is because it will cover a broader range of RSS and more geographically dispersed MSs.

In order that the clustering may be relevant for different antenna configurations, for example, it can be generated by using different transmission powers, different tilt angles

or a semi-smart antenna. That means modelling of the RSS distribution and radio coverage in each cluster is undertaken. For complex changes e.g. tilt, the reflections will be different, so the modelling is intended to capture the changes at least at a statistical level, such as the probability that the RSS will exceed a threshold can be computed at different locations. Modelling at detailed ray tracing level is not being performed. In the recent years, fingerprinting has attracted attention to predict user location in indoor environment. The implementation of this technology is quite simple and effective for indoor environment positioning, because it normally only needs the measurement of RSS or other non-geometric parameters at some locations to form a database of location fingerprints, and then to find out the better match fingerprints in the database. There has not been a great deal of research into how to adopt a fingerprinting approach in outdoor location estimation systems, because of the difficulty of the large amounts of data that need to be processed and tested.

In this work, a similar idea to a fingerprint-based positioning system is adapted for the outdoor environment. Firstly, a subset of transmitters is selected in the area of interest using principal component analysis (PCA) to overcome the drawbacks of transmitter selection methods. This initial selection is a coarse filter to clearly eliminate no useful BSs. A more refined selection occurs for each cluster. Then the deviation of the raw RSS from the estimated path loss model from the BS is clustered and the region is partitioned into several similar areas in terms of the effect of topography on the RSS. After clustering, the proposed approach returns to using the raw RSS in each determined cluster and use PCA again to rotate these raw RSS to independent principal components (PCs) within each cluster, which allows us to build a RSS distribution model within each cluster to support further positioning.

This chapter summarizes the theoretical fundamentals of the proposed measurement system in outdoor location estimation. Section 4.2 gives a brief outline of the proposed run-time positioning mechanism. The proposed transmitter selection method is presented in section 4.3. In section 4.4, the clustering scheme, which is used for partitioning

of the wireless environment, such as cell sectors, into regions that allow for accurate modelling of the propagation environment and prediction of the users' RSS distribution, is introduced in detail. Section 4.5 gives more details about how to transfer the raw RSS into independent ones within each cluster. The proposed clustering scheme is tested by using both simulated and real data collected from the outdoor environment and the evaluation of the simulation results are presented in section 4.6. Finally, section 4.7 makes a conclusion of this chapter.

4.2 The Overview of Outdoor Localisation System

The proposed positioning mechanism involves two phases: a *training phase* and an *online localisation phase*, and this is illustrated in Figure 4.1. This section focuses on the location estimation in a constant environment. Positioning in a dynamic environment will be discussed in chapter 7.

Finding the clustering scheme and creating the accurate RSS distribution models are

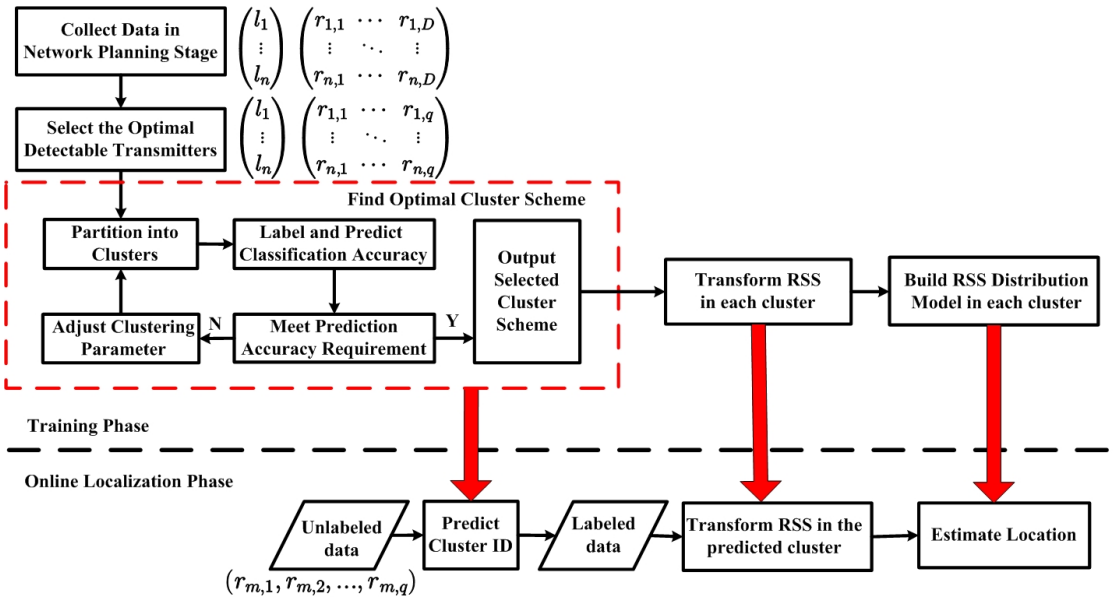


Figure 4.1: The overview of the proposed positioning mechanism

the main aims in the training phase. For this thesis, it can be considered the training is done once and for all. Before clustering the RSS data points, the first step is to choose the most representative subset of transmitters in the area of interest. The PCA is used globally to remove the least useful transmitters during the training phase to avoid unnecessary calculations. A clustering scheme is described to partition the environment into different disjoint regions where each region in RSS space maps to locations in the real environment that have similar RSS, as shown in the red dashed box in Figure 4.1. The black arrows in this figure show direction of flow from one step to another and the red arrows mean the needed information is provided by the results obtained in the training phase.

Once these models are constructed, they are applied to a new set of RSS values for online real-time location estimation. In the online phase, a new mobile phone user comes into the test-bed and he/she asks for positioning. That is to say, when a new MS with observed RSS tuple from nearby BSs has been collected, the better matching cluster for each new MS is found according to the received power using the K-Nearest Neighbour-Venn Probability Machine (KNN-VPM). Then its relative location in that cluster can be estimated. In this way, the algorithm is tolerant to a location calculation error.

This is achieved by creating clusters in the space of deviations of the observed RSS (in the training data) from the estimated log-distance path loss from each transmitter in the best chosen transmitter subset. Deviations of RSS are used for clustering rather than raw RSS, as this can approximate the decay with distance that can otherwise dominate the clustering. Since the clustering is performed in RSS space, the created clusters do not necessarily correspond to physically contiguous regions. For example, in a complex environment, two mobile users within the same cluster can be scattered over geographically dispersed locations or even be interspersed by the other users that belong to another different cluster. After partitioning using the number of clusters, the RSS distribution model is built in each cluster using PCA again. For each partition, the raw RSS is rotated into orthogonal dimensions and used to build RSS distribution models

for location estimation in the online phase.

4.3 Detectable Transmitters Selection

In a real environment, a mobile user can receive signals from many detectable transmitters within the area of interest. For example, for one of the test-beds (the Queen Mary campus in section 4.6.2.1), 29 BSs for a particular operator are detectable. In the training stage, assume a set of n MSs: the MS geographic location and the RSS measurements from all D neighbouring transmitters are collected. If one MS does not receive measurable signal strength from one typical transmitter, this means a null value indicator. A default value -120 dBm is given to this null value indicator, and also -120 dBm is the minimum strength of the signal strength received in the environment. The detectable transmitter selection process can be divided into two filtering steps:

Step 1: Neglect the transmitters that are far away from the target experimental area. If the number of MSs that cannot receive signals from transmitter j is less than $n/2$, transmitter j is ignored. Hence, the range of the detectable transmitters can be narrowed from D to $D^{'}$. So the RSS measurements received by all the training data from $D^{'}$ transmitters are described as

$$R = \begin{pmatrix} \vec{r}_1 \\ \vdots \\ \vec{r}_n \end{pmatrix} = \begin{pmatrix} r_{1,1} & \cdots & r_{1,D'} \\ \vdots & \ddots & \vdots \\ r_{n,1} & \cdots & r_{n,D'} \end{pmatrix} \equiv \begin{pmatrix} \vec{t}_1 \\ \vdots \\ \vec{t}_{D'} \end{pmatrix}^T \tag{4.1}$$

Where T is the symbol of matrix transpose. To better understand the equation (4.1), Figure 4.2 takes an example. Here $\vec{r}_i$ is a $D^{'}$-dimension row vector of RSS received by MS i from $D^{'}$ transmitters, while $\vec{t}_j$ is the n-dimension column vector of RSS received by all the n MSs from transmitter j, i.e. a different viewpoint on the same data set.

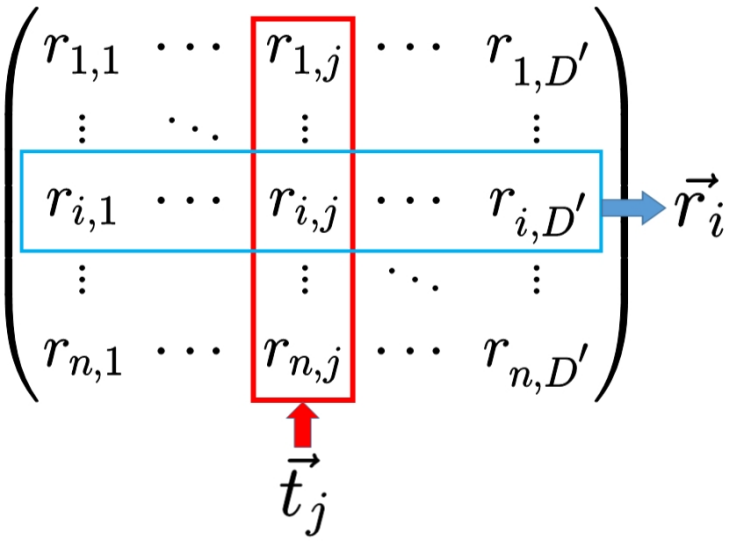


Figure 4.2: The explanation of equation (4.1)

Step 2: Use the PCA technique globally to project the measured RSS into a transformed signal space. The basis in the transformed space can be viewed as the linear combination of each transmitter with different weights (a.k.a principle components (PCs)), which represent the different contributions of each transmitter. It can be seen how many principal components are needed to express the percentage of the variability in the data set, and use these as the reduced dimensions (e.g. q dimensions $q < D^{'}$) with the added advantage that they are orthogonal.

Principal component analysis (PCA) is a statistical technique that uses a linear orthogonal transformation to convert a set of observations of possibly correlated variables into a set of uncorrelated variables called principal components (PCs). Mathematically, the first PC is a line passing through the multidimensional RSS mean and minimizes the sum of squares of the distances of the points from the line. So this axis has a large variability when measured along the axis (not orthogonal to it). The second PC is similar but constrained to be orthogonal to the first PC. The computed eigenvalue for each PC is proportional the sum of the squared distances of the points from their multidimensional mean (along the PC axis) and is often referred to as the "variance" of the PC. The sum

of all the eigenvalues is equal to the sum of the squared distances of the points from their multidimensional mean. PCA rotates the set of points around their mean in order to align with the PCs and this moves as much of the variance as possible into the first few orthogonal dimensions. Therefore, by removing the PCs that contribute little to the total variance, the aim is to project the entire data set to a lower dimensional space, but retain most of the information. This step can be further divided into four sub-steps as follow:

1. Calculate the matrix $\bar{R}$, each of its row vector is the mean value $(\bar{t}_1, ..., \bar{t}_{D'})$ of the training RSS data points in R from each transmitter, and then the $D' \times D'$ covariance matrix Σ of the training RSS data points can be obtained.

2. Calculate the eigenvalues and eigenvectors of Σ. The eigenvalues contains the variances for the PCs and the eigenvectors contains the linear coefficients for the principal components. Assume the eigenvalue $\{\lambda_1, ..., \lambda_{D'}\}$ is in descending order, and $\vec{e}_i$ represents the normalized eigenvector associated with λ_i. Thus, the principal component coefficients can be defined as $A = [\vec{e}_1, ..., \vec{e}_{D'}]$. So far, the principal components analysis itself has been accomplished.

3. Hence, to put the PCA to use, it needs to know what proportion each principal component represents of total variance, which can be expressed as

$$\omega_i = \frac{\lambda_i}{\sum_{i=1}^{D'} \lambda_i} \tag{4.2}$$

According to (4.2), it can remove the PCs that contribute little to the variance, and project the entire data set to a lower dimensional space, but retain most of the information. Figure 4.3 gives an example of how to choose the optimal PCs in the Queen Mary campus scenario. It shows how much variance in the data set is explained by which PC (by the bars shown in Figure 4.3) and how much variance is explained by the first 6 PCs (as seen by the blue line shown in Figure 4.3). It can be observed that the 6 first PCs can capture most of the variability in the signal

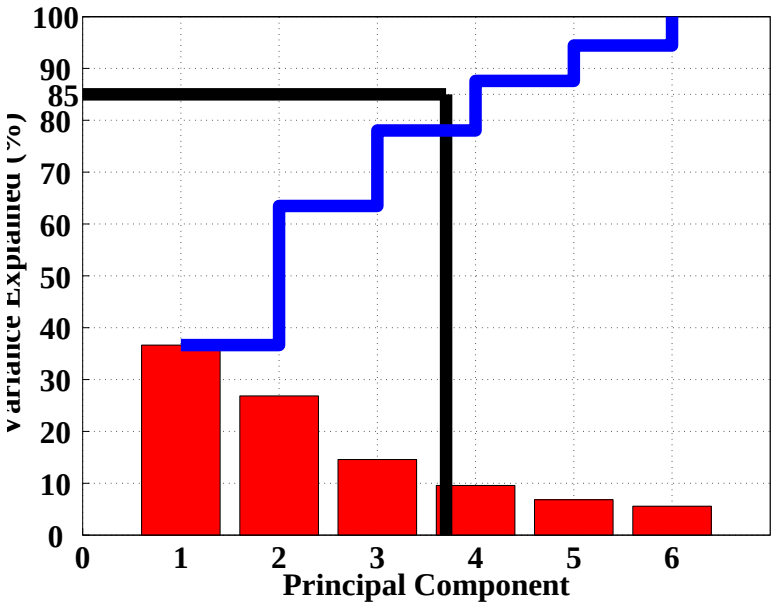


Figure 4.3: The cumulative variance accounted for by successive PCs in Queen Mary Scenario

strengths from the 29 BSs in Queen Mary Scenario and the first PC can expresses 36.4% of the total variance in the data set. Since the principal components are orthogonal, the amount of total variance expressed by the first 4 PCs is 87.60%, the sum of the proportions explained by them individually. So if it is needed to express 85% of the variability in the data set, it can be found that 4 PCs can be used as the reduced dimensions (e.g. $q \leq D$ dimensions), with the added advantage that they are orthogonal and give us almost 90% of the information about RSSs from the 29 BSs. In other words, the global PC can reduce the dimension by $25/29 = 86\%$ in this scenario. Hence, analyzing fewer pieces of information can give us almost the same results as analyzing the whole set of variables. In this way, the reduced dimensions can be denoted as q, which is 4 here. The optimized principal component is $A_{opt} = [\vec{e}_1, ..., \vec{e}_q]$.

4. Calculate component loadings of each transmitter to the q largest PCs.

$$u_{ij} = \sqrt{\lambda_i} e_{ij} (i \in [1, q], j \in [1, n]) \tag{4.3}$$

Here u_{ij} is the component loading of the j-th transmitter on the i-th PC, and e_{ij} is the j-th element of $\vec{e_i}$. Within each PC, one transmitter with the largest absolute value of component loading is chosen.

The use of PCA for selecting the most representative transmitters is also feasible in indoor environments, especially for small-/medium- size. For a large-size environment, e.g. one of the test-beds London Stratford Westfield shopping mall, it is not appropriate to apply PCA to choose the best subset of APs/BSs for the whole area. Details can be found in chapter 9 section 9.5.1.

4.4 The Proposed Clustering Scheme

Clusters are used instead of a uniform grid for a better model of complex topography as grid boundaries and topographic features do not necessarily align (this has been verified in direct comparisons on the data sets described here). Previous clustering localisation research [31] [52] [82] [83] does not pay attention to the cluster stability and to managing scalability issues without losing important correlation information. In the proposed clustering scheme, the Affinity Propagation method [84] is used for clustering and the Venn Probability Machine (VPM) [85] method is used to predict the probability of cluster membership and manage the trade-off between the estimation accuracy of cluster identification and the number of clusters to select the better clustering scheme. It not only pays attention to the cluster stability and to managing scalability issues without losing important correlation information, but also adjusts for RSS values to decrease the path loss effect that has the benefit, which the clusters are invariant to the power at the BSs or RSs, which will be explained later.

4.4.1 The Introduction of Affinity Propagation

Affinity Propagation [84] [86] clustering algorithm is a service-oriented architecture clustering method, which has been shown to produce clusters in much less time, and with much less error than traditional clustering techniques, such as K-means clustering, in [31]. For instance, K-means clustering method aims to partition a set of data into k clusters in which each data point belongs to the cluster with the nearest mean value. Although the main idea of K-means clustering is quite simple, it needs to determine the number of cluster (the value of k) in advance, and consumes too much processing time until the best results are selected. Due to those harsh requirements, it could not be a feasible method for a dynamic large wireless environment. Affinity Propagation clustering can solve the above issues. Affinity Propagation clustering can be utilized to identify a relatively small number of cluster centres (a.k.a exemplars) to represent all the points in a data set. In Affinity Propagation clustering, each data point can be viewed as a node in a network and simultaneously considered as a potential exemplar at first, and then real-valued messages are recursively transmitted along the edges of the network until a good set of exemplars and corresponding clusters emerges.

A review of the mathematical model of the Affinity Propagation approach is given below.

• **Exemplars**

Exemplars are the data points that are chosen to be the cluster centres. They are representative of themselves and some other data points that belong to the same clusters with them.

• **Input Arguments : Similarity and Preference**

Similarity

Affinity Propagation takes an input function of similarities, $s(i,k)$, where $s(i,k)$ in-

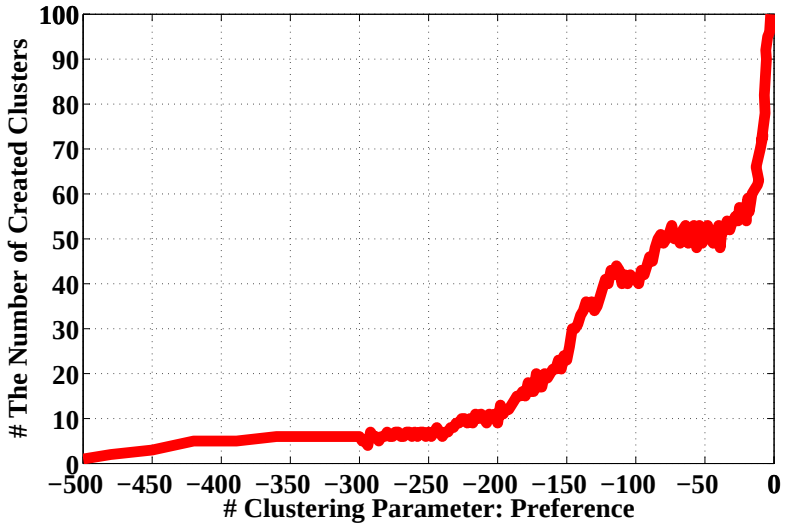


Figure 4.4: The effect of the preference value on the number of generated clusters

dicates how well suited data point k is to be the exemplar (a.k.a cluster centre) of data point i. If the data are real-valued, a common choice of similarity function could be the negative Euclidean distance between data points that a maximum similarity corresponds to the closest data points. Affinity Propagation clustering can be applied to use general notion of similarity, and the similarities can be positive or negative depending on the way in which the definition of similarity is appropriate for the application.

Preference

Each data point i has a self-similarity, $s(i, i)$, which reflects the prior suitability of data point i to be an exemplar and influences the number of exemplars that are identified. The self-similarity is also called "preference" that is another input parameter for this algorithm. Assigning a data point to a larger or smaller preference (self-similarity) value will respectively increase or decrease the possibility of the data point becoming an exemplar. In the beginning, Affinity Propagation clustering considers all data points as potential exemplars. So if one wants to make sure all data points are equally suitable as exemplars and there is no inclination toward particular ones as exemplars, the preferences of all data points should be set to the same value. In addition, the preference value can control the number of clusters that are generated. Figure 4.4 shows the relationship of

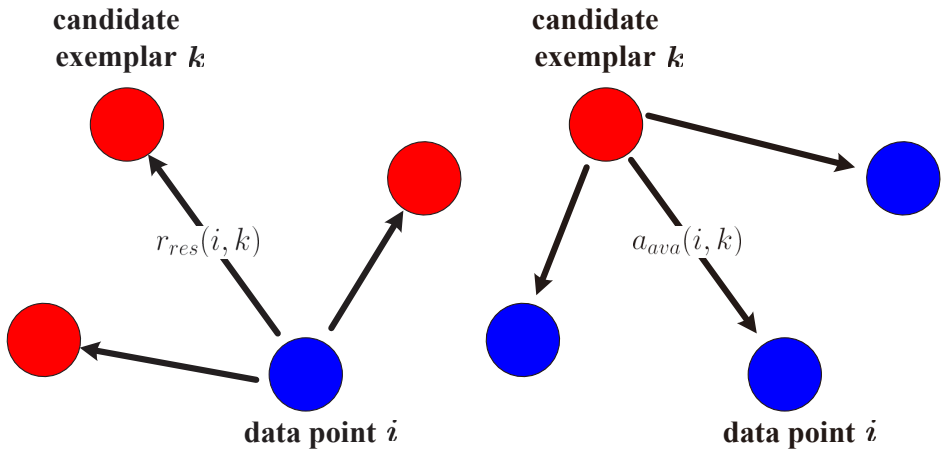
preference value on the number of produced exemplars. It can be clearly seen that low values of preferences will lead to a small number of clusters, while high values will find a large number of clusters that is produced.

• **Two Types of Message Passing : Responsibility and Availability**

In Affinity Propagation clustering method, two kinds of messages are exchanged between data points: *Responsibility message* and *Availability message*.

Responsibility

The responsibility message, $r_{res}(i,k)$, is sent from data point i to candidate exemplar data point k. A non-exemplar data point i informs each candidate exemplar whether it is suitable for joining as a member, as shown in Figure 4.5 (a). The message $r_{res}(i,k)$ indicates how well suited data point k is to be data point i's exemplar, taking into account competing other potential exemplars.



(a) Responsibility message (b) Availability message

Figure 4.5: Responsibility message and availability message

Availability

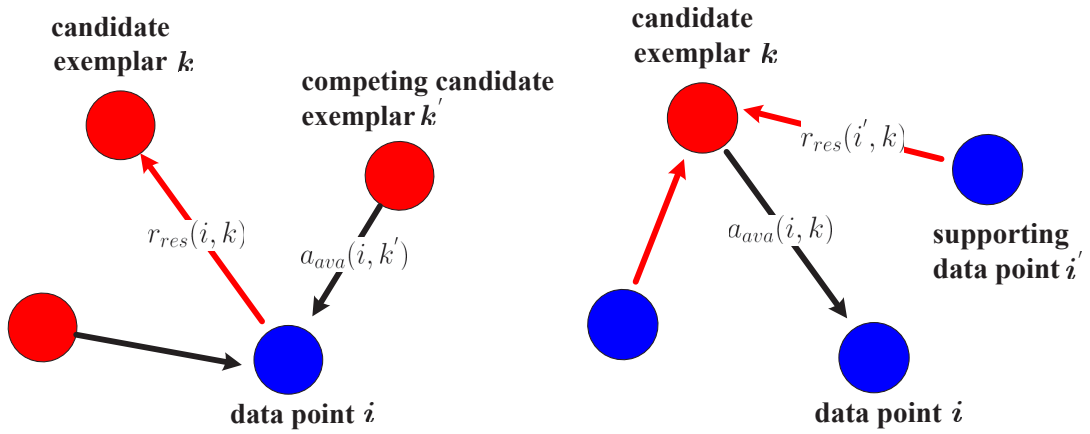
The availability message, $a_{ava}(i,k)$, is sent from candidate exemplar k back to potential cluster member data point i. A candidate exemplar data point k informs other

data points whether it is a good exemplar, as shown in Figure 4.5 (b). The message $a_{ava}(i,k)$, shows how proper it would be for point i to choose point k as its exemplar based on supporting the feedback from other data points. The calculation of availability only considers the positive responsibility messages from surrounding data points. If point k receives strong responsibility messages from surrounding data points, it will send a stronger availability message to indicate the suitable degree for it to become an exemplar.

It should be noted that the self-responsibility, $r_{res}(k,k)$ and self-availability, $a_{ava}(k,k)$ are two additional messages calculated for each data point k. Both of these two messages give accumulated evidence that point k is to be an exemplar, and are used to find the clusters. The self-responsibility message is based on the input preference value and the maximum value of availability message received from surrounding data points. It reflects how ill-suited it is to be assigned to another exemplar. In contrast to the self-availability message, the suitability for point k to be an exemplar is based on the number of the positive received responsibilities messages and their values.

In Affinity Propagation clustering, all data points are taken as potential exemplars simultaneously. In other words, all data points can be thought to be either candidate exemplars or cluster members, depending on whether they are sending or receiving responsibility or availability messages. Data points exchange and update these two messages in the network until a high-quality set of exemplars and corresponding clusters emerge. So there are some updated formulas for responsibility (4.4), self-responsibility (4.5), availability (4.6) and self-availability (4.7) should be complied with, and Figure 4.6 depicts how these two messages are exchanged between the data points. The algorithm begins by calculating the responsibilities with the availabilities set to 0.

$$r_{res}(i,k) = s(i,k) - \max_{k':k' \neq k}\{a_{ava}(i,k^{'}) + s(i,k^{'})\} \tag{4.4}$$



(a) Update responsibility message (b) Update availability message

Figure 4.6: Update responsibility message and availability message

$$r_{res}(k,k) = s(k,k) - \max_{k':k' \neq k}\{s(k,k')\} \tag{4.5}$$

$$a_{ava}(i,k) = \min\{0, r_{res}(k,k) + \sum_{i':i' \neq \{i,k\}} \max\{0, r_{res}(i',k)\}\} \tag{4.6}$$

$$a_{ava}(k,k) = \sum_{i':i' \neq k} \max\{0, r_{res}(i',k)\} \tag{4.7}$$

In the whole process, these two messages should follow the four equations above. While computing responsibilities and availabilities according to these simple updating rules will result in oscillations that are caused by "overshooting" the solution, so the responsibilities and availability messages are "damped" according to the following equation:

$$R_{res}(t+1) = (1-\lambda)R_{res}(t) + \lambda R_{res}(t-1) \tag{4.8}$$

$$A_{ava}(t+1) = (1-\lambda)A_{ava}(t) + \lambda A_{ava}(t-1) \tag{4.9}$$

Where $R_{res} = [r_{res}(i,k)]$ and $A_{ava} = [a_{ava}(i,k)]$ represent the responsibility matrix

and availability matrix respectively. t indicates the iteration times and λ is the damping factor that is used to avoid numerical oscillations. Because loopy belief propagation on which Affinity Propagation is based can be viewed as a particular kind of over-relaxation [87]. A damping factor is commonly needed in over-relaxation methods and it prevents the availability and responsibility updates from overshooting the solution and leading to oscillations in Affinity Propagation clustering. As long as the Affinity Propagation converges, the exact damping level should not have a significant effect on the resulting net similarity. In [84], they state that the damping factor λ should be at least 0.5 and less than 1, and the authors recommend setting the damping factor to 0.9. Higher damping factor λ will lead to slower convergence. If the Affinity Propagation does not converge, λ can be increased. But if the damping factor λ goes beyond 0.99, numerical precision issues will arise. For the experiments in this thesis, the value of λ is set to be 0.9 and there is no big difference when the damping factor falls within the range of $(0.5, 1)$.

• The Decisions of Clusters

The responsibilities and availabilities are messages that provide evidence for whether or not each data point should be an exemplar and if not, which exemplar that data point should be assigned. After the messages have converged, there are two ways to identify exemplars [87]:

1. For data point i, if $a_{ava}(i,i) + r_{res}(i,i) > 0$, then data point i is an exemplar.
2. For data point i, if $a_{ava}(i,i) + r_{res}(i,i) > a_{ava}(i,j) + r_{res}(i,j)$ for all j not equal to i, then data point i is an exemplar

This clustering procedure may be performed at any iteration of the algorithm, but final clustering decisions should be made once the algorithm stabilizes. The algorithm can be terminated once exemplar decisions become constant for some number of iterations, indicating that the algorithm has converged.

• The Process of Affinity Propagation

To begin with, the availabilities are initialized to zero for the first iteration: $a_{ava}(i,k) = 0$. So $r_{res}(i,k)$ is set to $(1-\lambda)$ times of the value of the difference of: the input similarity between data point i and data point k minus the largest competing similarity between point i and other competing potential exemplars (e.g. data point $k^{'}$ in Figure 4.6 (a)), which can be given by:

$$r_{res}(i,k) = (1-\lambda)(s(i,k) - \max_{k':k' \neq k} s(i,k^{'})) \tag{4.10}$$

After later iterations when some data points are assigned to other exemplars, their availabilities will drop below zero. This decreases the value of the corresponding similarity to which it is added and gradually removes them from the competition to be an exemplar. According to (4.6), the important part of availability update rule is its prescribed updated value. It is set to the self-responsibility $r_{res}(k,k)$ plus the sum of positive responsibilities candidate exemplar k receives from other points (e.g. the data point $i^{'}$ in Figure 4.6 (b), but it does not include the message destination: data point i). Only the positive portions of incoming responsibilities are added, because it is only necessary for a good exemplar to explain some data points well. If $r_{res}(k,k)$ is negative, it indicates that data point k is currently better suited as belonging to another exemplar rather than being an exemplar itself. The availability of point k as an exemplar can be increased if some other points have positive responsibilities for point k being their exemplar. For any data point during Affinity Propagation clustering, responsibilities and availabilities can be combined to identify exemplars.

4.4.2 Clustering Mobile Stations' RSS feedback

In the context of wireless networks, there are two benefits of Affinity Propagation clustering technique that can be applied to this research work: (a) the clusters emerge

naturally, rather than by specifying the number of clusters in advance and the number of clusters is related to a chosen "preference" value. (b) It allows for great flexibility in the face of dynamic environments, since all clustering parameters can be varied across iterations. Accordingly, the fundamental idea of Affinity Propagation clustering to partition the wireless environment based on the deviations RSS from the log-distance path loss models is used.

Assume there is a set of n MSs collected during training stage. For each MS, its geographic location and the RSS measurements from neighbouring transmitters are known. Let $\vec{r_i} = (r_{i,1}, r_{i,2}, ..., r_{i,q})$ represent the set of RSS from MS i from q antennas, i.e. BSs and RSs, in the area of interest, and the deviations from the q RSS log-distance path loss models create a tuple $\vec{\rho_i} = (\rho_{i,1}, \rho_{i,2}, ..., \rho_{i,q})$. In this work, clustering is based on the Mahalanobis distance rather than the Euclidean distance in signal space to create distinct and stable clusters. Because Mahalanobis distance function can avoid giving too much weight to correlated RSS values in the distance function and enables both non-linear and linear decision boundaries. Direct comparisons of the Euclidean distance and Mahalanobis distance on real data set are given in chapter 5 section 5.5 later. For any two MSs, such as MS i and MS k, the similarity between them can be expressed as:

$$s(i,k) = -\sqrt{(\vec{\rho_i} - \vec{\rho_k})^T \Sigma^{-1} (\vec{\rho_i} - \vec{\rho_k})}, \forall k \neq i \tag{4.11}$$

Because the signal strength received by MSs from different BSs can be correlated, the covariance matrix Σ in signal space used in (4.11) is to describe the mutual dependence of the signal strength received by any two MSs from different BSs. If there are q BSs nearby, Σ can be estimated as:

$$\Sigma_{q \times q} = \begin{bmatrix} \Sigma_{1,1} & \cdots & \Sigma_{1,q} \\ \vdots & \ddots & \vdots \\ \Sigma_{q,1} & \cdots & \Sigma_{q,q} \end{bmatrix} \tag{4.12}$$

Where $\Sigma_{j,p} = \frac{1}{(n-1)} \sum_{i=1}^{n} (\rho_{i,j} - \bar{\rho}_j)(\rho_{i,p} - \bar{\rho}_j)^T$, $1 \leq j, p \leq q$, n is the number of RSS tuples, T is the symbol of the vector $(\rho_{i,j} - \bar{\rho}_j)$ transpose and $\bar{\rho}_j$ is the average deviation RSS value of MSs from BS j, $\bar{\rho}_j = \frac{1}{n} \sum_{i=1}^{n} \rho_{i,j}$. If the covariance is not used and it is assumed that RSS is independence as in the Euclidean distance, then highly correlated variables are given too much weight in the distance function.

The use of deviations can eliminate to some extent the effects of distance dependent path loss attenuation, and so better capture the effects of multipath and shadowing, which mainly depends on the topography. Using the raw RSS leads to clusters where the similarity is dominated by the distance path loss, which is approximated anyway by the estimated path loss model. Direct comparisons on different data sets demonstrate this and they are given in chapter 5. The deviations from the log-distance path loss model are obtained based on the RSS data during the training stage. According to the log-distance path model [11], the RSS measurement P_{rss} (in dBm) at the distance d from the transmitter can be formulated as in the equation below:

$$P_{rss} - PTR = \kappa + \gamma log(d/d_0) \tag{4.13}$$

Where PTR represents the transmit power of the transmitter, d_0 is reference distance for the antenna area and the value of it is set to 100 meters in this research. The values of parameter γ and κ are heavily dependent on the environment and are estimated by a least squares linear regression model (4.13) from training data for the transmitter.

Figure 4.7 shows the process of how to cluster MSs RSS data using Affinity Propagation clustering algorithm in this research. Each box represents a step respectively corresponding to step 1 to step 6. Step 3, step 4 and step 5 are iterated for several times or until convergence. Here are the details of each step as below.

Step 1: Collect current mobile users' feedback, analyse and process current RSS data.

Step 2: Construct similarity function according to (4.11) and define preference value that may influence the number of clusters that are created.

Step 3: Update and compute the responsibilities messages according to (4.4) and (4.5). The responsibility update rule let all potential exemplars compete for ownership of a MS.

Step 4: Update and compute the new availabilities messages according to (4.6) and (4.7). The availabilities update to gather the evidence whether each candidate exemplar would make a good exemplar or not.

Step 5: Calculate the exemplars for all MSs using the result of iteration of step 3 and step 4. Combine responsibilities and availabilities to monitor the exemplars decisions. For any MS i, if $r_{res}(i,i) + a_{ava}(i,i) > 0$, then identify MS i to be the current estimated exemplar.

Step 6: If the estimated exemplars produced in step 5 stay unchanged for a certain number of iterations or the number of iterations reaches the maximum value, the program will assign other MSs to the exemplars depending on which one of exemplars is most similar to them, and then output cluster result. Otherwise go back to step 3.

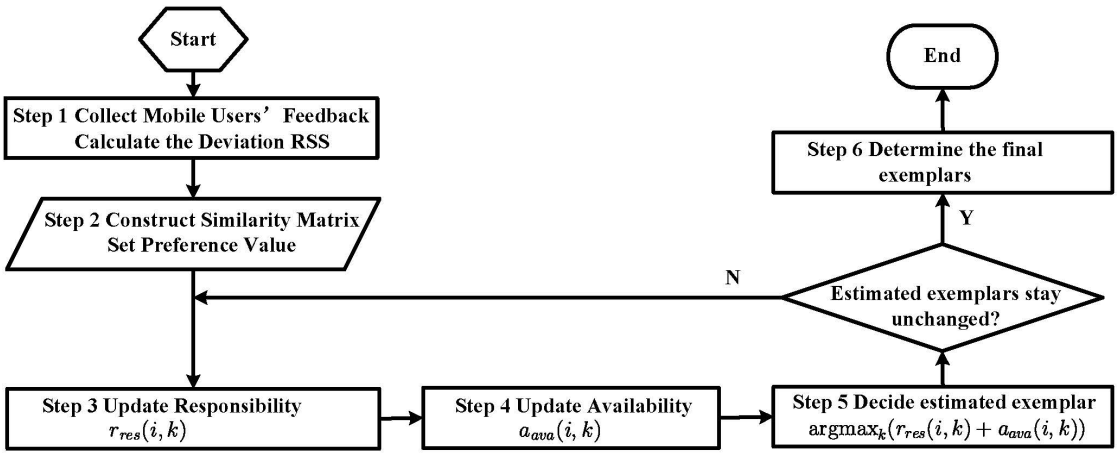


Figure 4.7: Flowchart of Affinity Propagation clustering

4.4.3 Estimation of the Accuracy of Cluster Identification

The **Venn Probability Machine** (VPM) learning technique [85] is a classification system usually applied on top of an existing learning algorithm, e.g. KNN, to augment predictions with probability estimates. In this work, the VPM is used to estimate the probability of cluster membership and to manage the trade-off between the estimation accuracy of cluster identification and the number of clusters. The RSS training data set are randomly split into two halves as the cluster training and cluster testing sets. The cluster training set is used as representatives of the clusters that have been obtained; and the cluster testing set is allocated to clusters based on KNN. This allows the computation of the probability that each RSS tuple in the cluster testing set belonging to each cluster, to find the most likely cluster ID for the testing data point and to validate the accuracy of cluster classification for the testing data point. The accuracy of the cluster identification prediction is also used to support the MSs clustering. According to the required cluster estimation accuracy, the preference value is adapted based on the calculated accuracy and the clustering number found.

Let R represent the space of RSS tuples for MSs from the neighbouring BSs, and C be the space of cluster IDs and $Z = R \times C$, which denotes the pair (RSS tuple, cluster ID) for every MS in the area of interest. The RSS tuple will be referred to as a data point. The clustering set $C = \{C_1, C_2, C_3, ..., C_T\}$ and T is the number of clusters. The training data set, TR, can be represented as $TR = \{z_1, z_2, z_3, ..., z_N\}$, where $z_n = [r_n, c_n]$,$c_n \in C$. The testing data set, TS, can be denoted as $TS = \{z_{N+1}, z_{N+2}, z_{N+3}, ..., z_{N+S}\}$. Suppose the cluster ID of every test data point is unknown, the objective is to assign the estimated cluster ID for every test data point using its RSS tuples and make an assessment of the cluster ID prediction accuracy by comparisons with the known correct cluster ID of these test data points. The process is described in Algorithm (4.1).

First, choose one test data point from the test data set (step 1) and combine it with the training set TR to form a new data set (step 2). Use KNN algorithm to obtain a list of

Algorithm 4.1 K-Nearest Neighbours Venn Probability Machine

Requried:

k_{max}: the maximum value of nearest neighbours used
Cluster ID: $\{C_1, C_2, C_3, ..., C_T\}$
Training data set $TR = \{z_1, z_2, z_3, ..., z_N\}$, $(z_n = [r_n, c_n], c_n \in C)$
Test data set $TS = \{z_{N+1}, z_{N+2}, z_{N+3}, ..., z_{N+S}\}$

Steps:

1: **for** $s = 1$ **to** S **do**
2: $TM = \{z_1, z_2, z_3, ..., z_N, z_{N+s}\}$.
3: Using RSS to calculate the distances and get each z_i its neighbours $Neighbour(z_i)$ in a descending order of respective distance.
4: **for** $t = 1$ **to** T **do**
5: Assign $z_{N+s} \in C_t$
6: **for** $k = k_{max}$ **to** 1 **do**
7: **if** $\exists z_p \in TR$ such that $Neighbour(z_p)(1:k) = Neighbour(z_{N+s})(1:k)$ **then**
8: $k_{eff} = k$
9: Put z_p into $\mathcal{Z}$
10: Fill $\mathcal{Z}$ with all other z_q in TR that satisfy
$Neighbour(z_q)(1:k_{eff}) = Neighbour(z_{N+s})(1:k_{eff})$
11: Break
12: **end if**
13: **end for**
14: **for** $\tau = 1$ **to** T **do**
15: Calculate the frequency of each cluster
$P_{t,\tau} = \frac{sizeof(\{z_\mu \in \mathcal{Z}, c_\mu \in C_t\})}{sizeof(\mathcal{Z})}$
16: **end for**
17: **end for**
18: $c_{N+s} = \text{agr max}_{c_{N+s} \leq T}(\min\{P_{c_{N+s},1}, ..., P_{c_{N+s},T}\} + \max\{P_{c_{N+s},1}, ..., P_{c_{N+s},T}\})$
19: **end for**

neighbours for each data point (step 3). Then the process works recursively on different cluster IDs (step 4). Specifically, the test data point is assigned with current cluster ID (step 5), so each data point gets a list of cluster IDs which can be converted from its list of neighbours. *Compare the first k (initially k_{max}) cluster ID in the respective list between the test data point and each training data point* (step 6). If there exists no training data points that has the same sequence of the first k cluster ID with the test data point (step 7), decrease k by one (step 6) and repeat. Once a training data point satisfying this condition is found, the effective value of k is set as k_{eff} (step 8). Then, put this training data point and all other eligible training data points into collection $\mathcal{Z}$ (step 9 and step 10). The normalised frequency of each cluster can be obtained by counting the number of MSs in $\mathcal{Z}$ (step 14 to step 16). These probabilities also compose the corresponding column of the frequency matrix. Repeat step 4 until all the cluster IDs

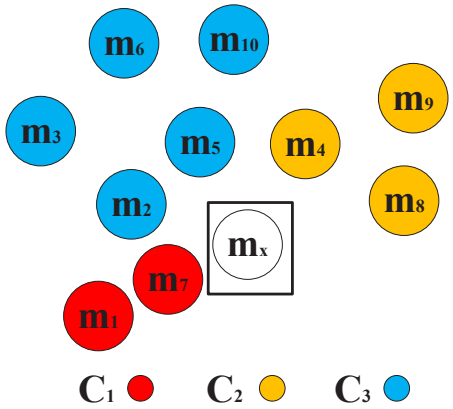
are analysed. As a result, all the columns of the matrix can also be filled. Finally, the mean of the maximum and minimum values of each row is regarded as the probability that the selected test data point belongs to each corresponding cluster. Therefore, the cluster of the test data point can be estimated as the one with the largest probability (step 18). The cluster ID of the other test data points can be estimated in the same way.

4.4.3.1 An example of Venn Probability Machine

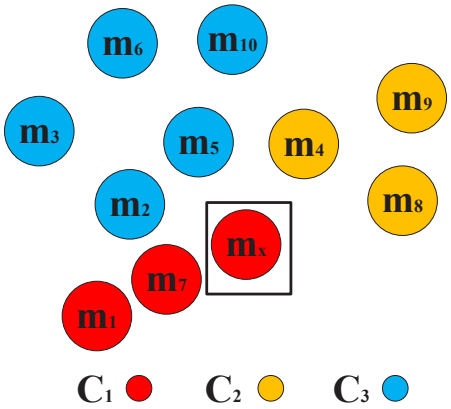
A simple example is given for a better understanding of how the VPM algorithm works. Suppose that the RSS data of ten MSs have been measured in the area of interest, where only four BSs are taken into account. Based on the Affinity Propagation clustering method, assume that three clusters are created which can be denoted as $\{C_1, C_2, C_3\}$, as shown in Figure 4.8 (a). Thus, the training set can be set as the RSS measurements of these ten MSs and their corresponding cluster IDs.

Now given a new MS m_x with RSS data, its cluster ID can be estimated as followed. Firstly, combine m_x with the training data set and then calculate the similarity between any two MSs based on their RSS. For each MS, K nearest neighbours can be estimated in a descending order (assuming K=3) as shown in Figure 4.9, in which the corresponding cluster IDs are also listed.

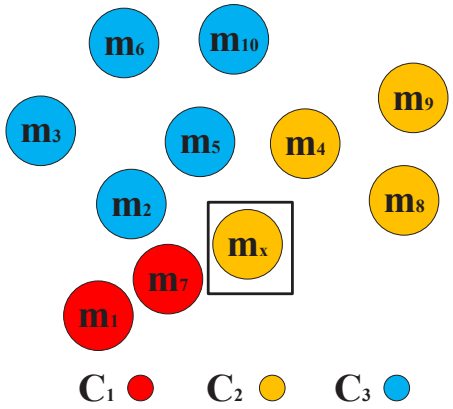
Since the m_x's cluster ID is unknown, it can be hypothetically assigned with all of the cluster IDs in turn in order to calculate the probability of m_x being in each cluster. Here m_x is firstly assigned to cluster C_1 (Figure 4.7 (b)). Consequently, the cluster ID list of each MS's three nearest neighbours need to be updated as shown in Figure 4.10. The next step is to search for the MSs which have the same cluster ID list with m_x's. However, as seen in Figure 4.10, there is no MS in the training data set that satisfies this condition. Therefore, these searching steps will be repeated with a smaller K, which turn to be two. Now, there are two MSs m_2 and m_{10} that have the same two-nearest-neighbour list with m_x can be selected. They are stored in a category $\mathfrak{Z}$,



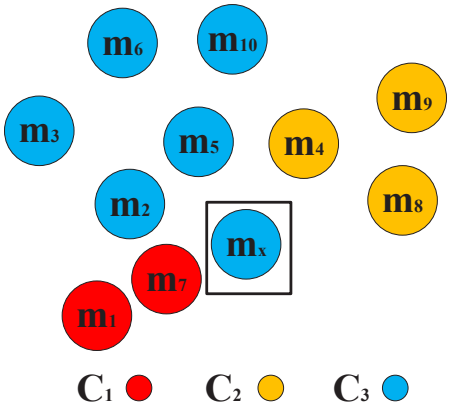
(a) Estimate cluster ID for MS m_x



(b) Assign to cluster C_1 hypothetically



(c) Assign to cluster C_2 hypothetically



(d) Assign to cluster C_3 hypothetically

Figure 4.8: Estimate cluster ID for MS m_x

so $\mathcal{Z} = \{m_2, m_{10}\}$. Then the normalised frequency of each cluster can be obtained by counting the number of MSs. For this example, the frequencies of all clusters are 0,0,1.

Similarly, cluster ID C_2 and C_3 are in turn assigned to m_x (Figure 4.8 (c) and (d)) to obtain the frequency of each cluster with the initial K value of three. The final frequencies under different assumptions of the m_x's cluster ID in this example is shown in Table 4.1. The average value of the maximum and minimum frequencies of each row is regarded as the probability that m_x belongs to each cluster. Finally, the cluster of m_x will be estimated as the one with the largest average frequency value, which is C_3 in this example.

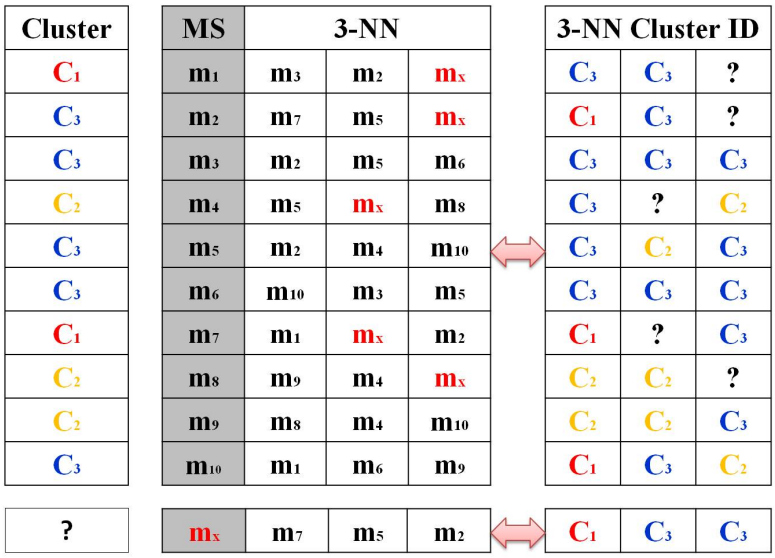


Figure 4.9: The list of three nearest neighbours of all the MSs

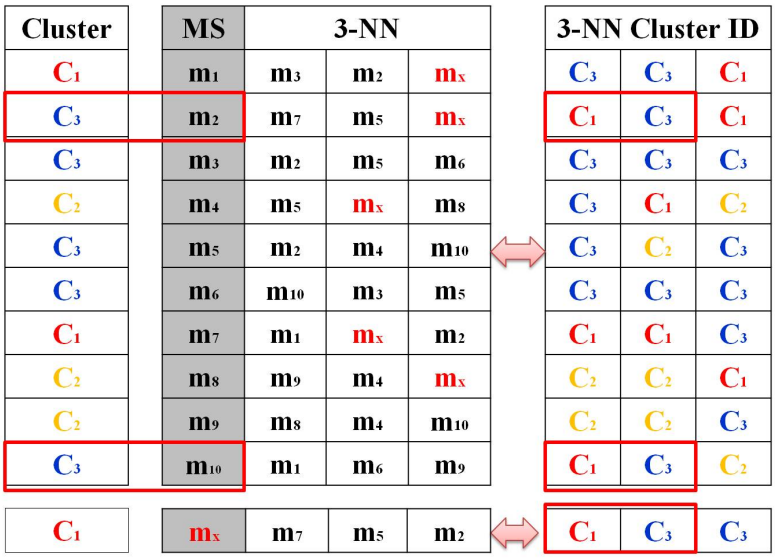


Figure 4.10: The list of three nearest neighbours when hypothetically assigns C_1 to m_x

Table 4.1: Frequencies Table for the MS m_x's Possible Cluster ID

Frequency \ Assumption	C_1	C_2	C_3	Normalised Frequency
C_1	0	0	0.5	$\frac{(0+0.5)}{2} = 0.25$
C_2	0	0	0	$\frac{(0+0)}{2} = 0$
C_3	1	1	0.5	$\frac{(0.5+1)}{2} = 0.75$

4.5 Selecting the Number of Clusters

One of the objectives is to have a coherent partitioning in the RSS space. A stable clustering can improve the estimation accuracy of cluster member assignment and hence can give useful information for the purpose of monitoring a dynamic MS environment and predicting users' locations. Accordingly, the issue of determining the number of clusters is taken in to account.

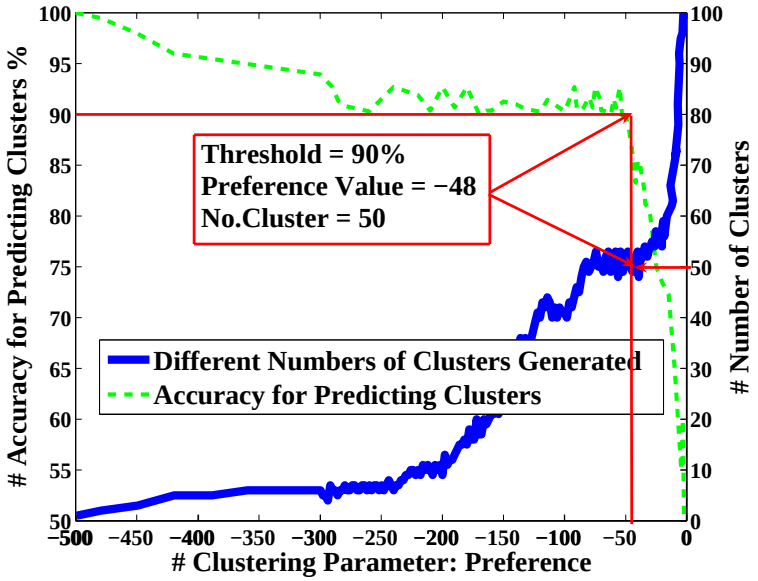


Figure 4.11: An example of the selection of cluster number

An example is given in Figure 4.11 to illustrate the identification of the number of clusters. The thick blue line represents the relationship between the cluster preference parameter value and the number of clusters generated. The green dashed line depicts the dependence of the cluster prediction accuracy on the number of clusters created. There is a trade-off between the number of generated clusters and location estimation. A greater number of clusters generated in the training period reduces the cluster prediction accuracy e.g. if there were one cluster, the cluster prediction accuracy would be 100%, but the location estimation would be poor), but results in a higher precision in location estimation, conditional on having chosen the correct cluster. Here, the objective is to

find the right balance between the accuracy of cluster identification and the number of clusters. The number of clusters is the maximum number of clusters that can still satisfy the accuracy requirements for cluster prediction. The cluster shape is determined by the location of training data set in this cluster. Seen from Figure 4.11, if the threshold accuracy of cluster identification is taken as 90%, the corresponding maximum number of clusters is 50. The training data set is collected first and the collection time depends on the size of area required. After cluster identification, models (such as other regression models) are fitted to each cluster. As the points are similar but not identical, a local and representative traffic distribution model can be determined from the points in the cluster.

Hence, using the clustering scheme, the terrain is divided into a set of clusters $C = \{C_1, C_2, ..., C_N\}$, where N is the total number of clusters. These created clusters construct a radio map, which not only capture the characteristics of the signal propagation in a given environment, but also avoid the modelling of the complex radio propagation and reduce the computational cost of coverage prediction. If M denotes the radio map, the i-th element in the radio map can be expressed as

$$M_i = (C_i, R_i, L_i), i = 1, ...N \tag{4.14}$$

Let n_i be the number of training MSs within cluster C_i. R_i denotes the RSS measurements in cluster C_i, which is given by

$$R_i = \begin{pmatrix} \vec{r}_1 \\ \vdots \\ \vec{r}_j \\ \vdots \\ \vec{r}_{n_i} \end{pmatrix} = \begin{pmatrix} r_{11} & \cdots & r_{1,b} & \cdots & r_{1,q} \\ \vdots & \ddots & & & \vdots \\ r_{j,1} & & r_{j,b} & & r_{j,q} \\ \vdots & & & \ddots & \vdots \\ r_{n_i,1} & \cdots & r_{n_i,b} & \cdots & r_{n_i,q} \end{pmatrix}_{n_i \times q} \equiv \begin{pmatrix} \vec{t}_1 \\ \vdots \\ \vec{t}_b \\ \vdots \\ \vec{t}_q \end{pmatrix}^T \tag{4.15}$$

Here $\vec{r}_j$ is a q-dimension row vector of RSS received by MS j from q antennas in

cluster C_i. $L_i = (l_{i1}, ..., l_{ij}, ..., l_{in_i})$ consists of the geographical locations, l_{ij} of MS j in cluster C_i. $\vec{t_b}$ is the n_i-dimension column vector of RSS received by all the n_i MSs from transmitter b, i.e. a different viewpoint on the same data set.

4.6 RSS Transformation within each cluster

When computing locations in the real environment, the correlation between signal strengths cannot be neglected. For example, for the real RSS data points collected from Queen Mary Campus given in section 4.6.2.1, Table 4.2 presents the RSS correlations from different BSs in one typical cluster. The RSS samples from different BSs can have correlations as high as 0.9, as shown in Table 4.2. The main challenge is to determine how to improve location estimates despite such high correlations.

Table 4.2: The correlation between signal strength in one typical cluster in Queen Mary Scenario

	BS 1	BS 2	BS 3	BS 4
BS 1	1	0.8293	0.7117	0.4387
BS 2	0.8293	1	0.9084	0.4656
BS 3	0.7117	0.9084	1	0.6687
BS 4	0.4387	0.4656	0.6687	1

For each cluster, the author uses PCA again to transform the correlated q-dimensional training RSS data set into a new basis, which are uncorrelated, and then builds the regression model for each cluster for each BS using the transformed uncorrelated RSS data points in that cluster. According to (4.14), the radio map in each cluster C_i is $M_i = (C_i, R_i, L_i)$. So for each cluster C_i:

Step 1 and 2: Calculate the $q \times q$ covariance matrix of the training RSS data points in this cluster and then compute the eigenvalues and eigenvectors of this covariance matrix to obtain the principal component coefficients A_i.

Step 3: Transform the original RSS data set of each cluster to obtain the new training RSS data set. Here each of the matrix $\bar{R}_i$'s row vector is the mean value $(\bar{t}_1, ..., \bar{t}_b, ..., \bar{t}_q)$

of the training RSS data points in R_i from each BS.

$$R_i^{'} = A_i^T \cdot (R_i - \bar{R}_i) \tag{4.16}$$

For convenience, one takes the row vector $\vec{r}_j^{'}$ and column vector $\vec{t}_b^{'}$ to represent $R_i^{'}$, which can be expressed as

$$R_i^{'} = \begin{pmatrix} r_{11}^{'} & \cdots & r_{1,b}^{'} & \cdots & r_{1,q}^{'} \\ \vdots & \ddots & & & \vdots \\ r_{j,1}^{'} & & r_{j,b}^{'} & & r_{j,q}^{'} \\ \vdots & & & \ddots & \vdots \\ r_{n_i,1}^{'} & \cdots & r_{n_i,b}^{'} & \cdots & r_{n_i,q}^{'} \end{pmatrix}_{n_i \times q} = \begin{pmatrix} \vec{r}_1^{'} \\ \vdots \\ \vec{r}_j^{'} \\ \vdots \\ \vec{r}_{n_i}^{'} \end{pmatrix} \equiv \begin{pmatrix} \vec{t}_1 \\ \vdots \\ \vec{t}_b \\ \vdots \\ \vec{t}_q \end{pmatrix}^T \tag{4.17}$$

In the following chapters, PCA is applied into each cluster, and then two different positioning algorithms are used to estimate user location, which will be described in chapters 5 and 6 respectively.

4.7 Experimental Results

In this section, the simulated and real data are tested to evaluate the performance of the proposed mechanism.

4.7.1 Results with Simulated Data

Two different sets of data are considered: one data set is obtained from a simplified urban environment propagation model and the other data set is generated by a network planning tool for the island of Jersey (primarily rural).

4.7.1.1 Outdoor Scenario 1: A Simple Simulated Urban Propagation Model

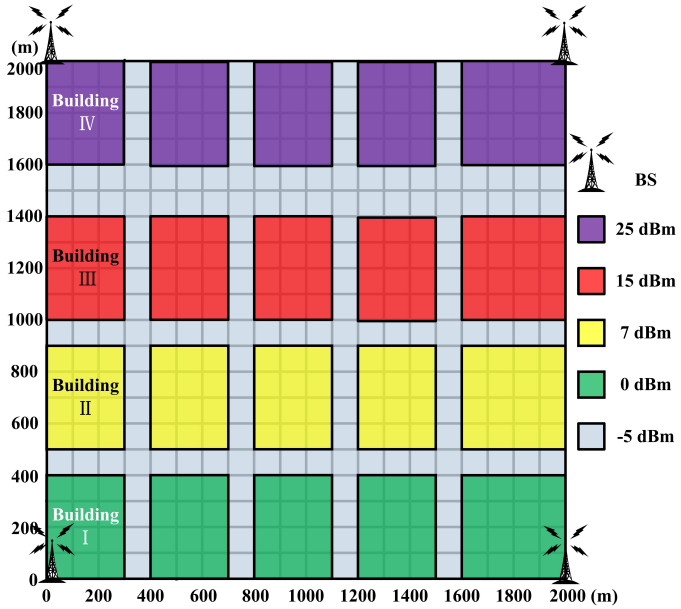


Figure 4.12: Topology of the simulated urban environment

Table 4.3: Configuration Parameters Used in the Simulation

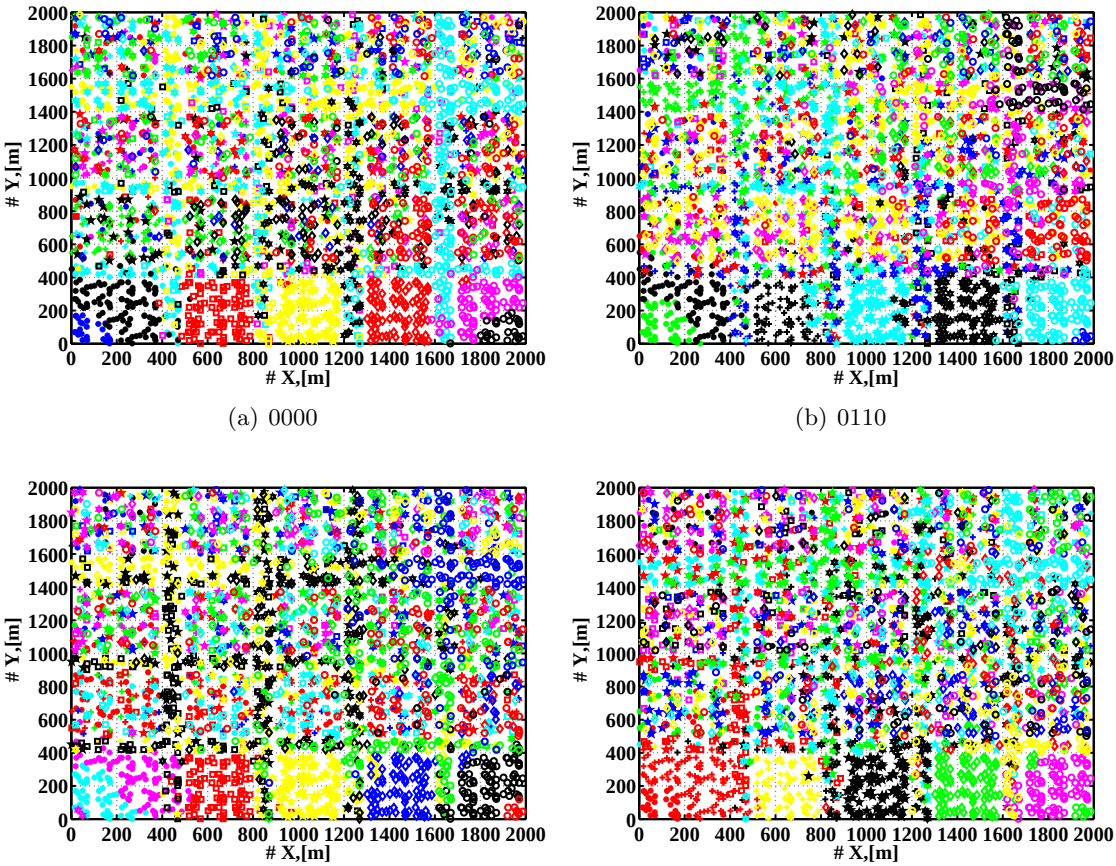
System Setting	
Distance between BS to BS	2.0 km
Minimum Mobile-to-BS distance	20 m
Total Number of MSs	3200 (8 MSs in each grid element)
Propagation Environment	
Minimum Transmit Power	40 dBm
Maximum Transmit Power	48 dBm
Shadowing Deviation	
Building I	0 dBm
Building II	7 dBm
Building III	15 dBm
Building IV	25 dBm
Street	-5 dBm

A 2 km $\times$ 2 km square area with four BSs at each corner is built as shown in Figure 4.12. The propagation model used in the simulation is based on the reference propagation model of COST-231 urban that is the combination of typical logarithmic path loss model and Rayleigh fading model. For simplicity, reflection, diffraction and scattering effects are not taken into account. The simulation area is divided into 20 $\times$ 20 elements by a

rectangular grid. For each grid element, there is a propagation feature that represents the shadowing variation in the corresponding grid element in the urban environment. The shadowing feature in each grid is given by the mean of the shadowing variation deviation of the uniform distribution of (0, 1). The mean of the shadowing variation is -5 dBm, 0 dBm, 7 dBm, 15 dBm and 25 dBm respectively, depending on the grid element which one of building blocks or street belongs to in Figure 4.12.

Furthermore, the MSs are uniformly distributed over the whole area and an equal number of sample MSs are selected from each grid element. Every MS in this area could receive signal strength from the four BSs. The important simulation configuration parameters are given in Table 4.3. To test whether the clusters represent the features of the topography, it needs the simulation model used to generate RSS from each BS to accommodate different physical situations. Suppose there are two settings for BS transmit powers: high power (48 dBm) and low power (40 dBm). There can be 16 possibilities from different combinations of powers for the four BSs. Let "1" denote the high power and "0" be the low power. These 2^4 factorial experiments can be simply expressed as: 0000, 0001,..., 1110, 1111.

In order to better analyse and compare with the results of 16 settings, all the MSs are chosen at exactly the same locations in each experiment that the RSS data set is collected. Figure 4.13 shows four examples of the results of clustering MSs based on deviation RSS over different powers at the four BSs. It can be seen from each sub-figure in Figure 4.13, different colours represent different clusters, and the cluster distribution can be seen to reflect the topological feature of the simulation area to some extent, especially for the places with a relatively small shadow variation. For the area with relatively large shadow variation, such as block IV in Figure 4.12, the cluster distribution is scattered with respect to the geographical locations of the MS, though in the four dimensional RSS space they are compact. The number of clusters produced in every test is not exactly the same but quite close, nearly 50 clusters. These results from all the 16 groups have shown that the distributions of created clusters all have the roughly same structure as



(a) 0000 (b) 0110

(c) 1101 (d) 1111

Figure 4.13: Clustering results in the outdoor scenario 1

would be expected mathematically as a result of using the deviations. Besides, within the same parameters, each experiment has been tested many times and the results showed good stability in the clustering. Although the simulated urban model used is simple, it can be further improved by analysing the azimuth and elevation power distribution of the transmission antenna to make sure whether this approach can be used in various scenarios in wireless networks.

4.7.1.2 Outdoor Scenario 2: The Island of Jersey area

The pilot signal strength data for the island of Jersey is obtained from network planning tool ASSET 3G. Figure 4.14 shows the topographic map of the centre of the island of Jersey with six BSs covering an area of 8 km $\times$ 6 km and the clustering result is depicted

Figure 4.14: The topography map in outdoor scenario 2

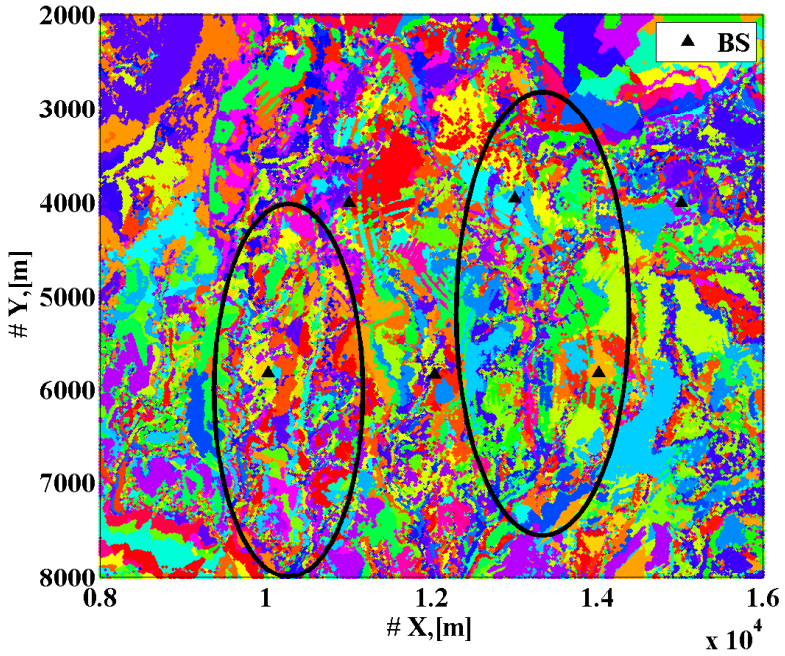


Figure 4.15: Clustering result in outdoor scenario 2

in Figure 4.15.

As seen from the results in Figure 4.15, the clusters can generally represent the features of the current geographical patterns to a certain extent, particularly the contour of highways and roads. These results are consistent with the results of scenario 1. The less shadowing variability in an area, the more topographical features are predicted by

clustering. Each experiment has been tested several times and the results demonstrated the performance stability. Because of the complex terrain of the central area, 160 different clusters are created in the central area. Given the large number of test points and the complexity of the model it seems feasible to adopt this approach and the RSS clusters have a mapping to the topography that is meaningful.

4.7.2 Results with Real Data

To test the proposed mechanisms in a real environment, two scenarios have been set up. One environment test-bed is around Queen Mary campus and the other is a three-day music festival in London Victoria Park. Both of these are essentially outdoors as they are based on GSM signals and depended on GPS to validate the accuracy outdoors. The RSS data of a GSM network was collected by a mobile app on an Android smart phone. In general, mobile phones are in communication with one or more BSs during and between connections. The mobile phone measures the received signal strength from nearby BSs and attempts to access the BS with the strongest signal when a connection is to be established. As the author uses the smart phone and moves around Queen Mary University or London Victoria Park, within every 1 second, the mobile application can record the exact latitude and longitude of the current location from GPS in the smart phone, and collect the varying signal strengths from the surrounding BSs. The locations of all the nearby BSs obtained from the server of Sony Ericsson are reported. More details about these two scenarios including the downloadable raw data can be found in [88].

4.7.2.1 Outdoor Scenario 3: Queen Mary campus

The Queen Mary campus covers a 475 m $\times$ 365m area and is in an urban area with tall buildings. In Figure 4.16, the colour-line area represents the result of clustering 9277 test points. In this case, the number of clusters is 70. 2 to 4 PCs, i.e. q=2 to 4, are

chosen for each cluster (depending on the cluster).

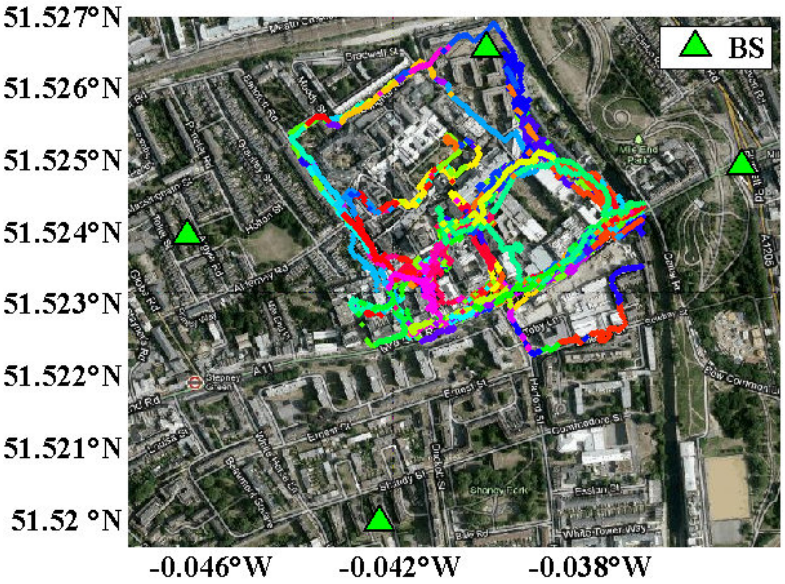


Figure 4.16: Clustering result in outdoor scenario 3

4.7.2.2 Outdoor Scenario 4: Three-day Music Festival in London Victoria Park

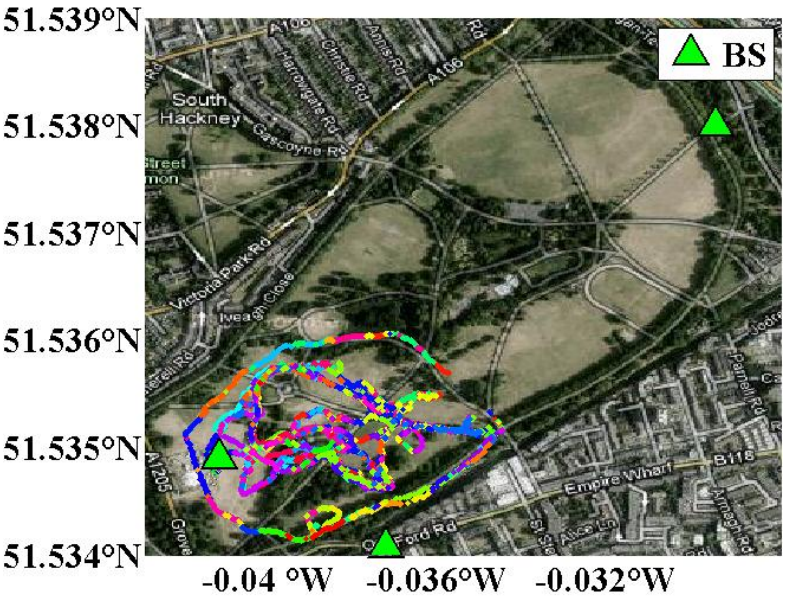


Figure 4.17: Clustering result in outdoor scenario 4

The GSM RSS data is collected from a three-day music festival held in London Victoria Park that covers a 450 m × 240 m area. There are only 10 BSs around the festival field, 3 BSs are chosen. The data sets are partitioned into three separate parts according to the day collected, that is, Day 1, Day 2 and Day 3. 2095 RSS samples collected on the first day are used as the training data set and then used to create clusters based on the RSS measurement. The data sets from Day 2 and Day 3 will be analysed later in chapter 7. As illustrated in Figure 4.17, the coloured-line area represents the result of clustering 2095 test points on Day 1. In this case, the number of clusters is 52.

4.8 Summary

In this chapter, the proposed positioning measurement mechanism was introduced, which can be applied into the real environment by monitoring users' RSS values continually. The simple global dimensionality reduction using PCs was used to select the most representative detectable transmitters. In addition, this proposed scheme clustered the RSS tuples based on deviations from an estimated RSS attenuation model and then transformed the raw RSS in each cluster into new uncorrelated dimensions, using PCs again.

In order to evaluate the feasibility of the clustering scheme, relevant experiments were carried out. According to the experiment results, it can be concluded that the performance of the proposed clustering scheme that contribute to the greater accuracy in the test-beds: a) the use of deviations from the observed path loss model for each RSS component rather than the raw RSS. This also results in the clusters being invariant to the BS/RS power; b) the accurate estimation of the cluster membership probability and the number of clusters to manage the trade-off between cluster size and accuracy of cluster modelling.

The limitations of the proposed clustering are: a) it will take a relatively long time to generate the number of clustering scheme during the training phase; b) for the clus-

tering scheme, the aim is to combine with VPM to find the number of clusters in the target area, no matter what kind of clustering method used, e.g. K-means and Affinity Propagation methods. Therefore, the difference between different clustering methods in a real environment will be compared in future work.

The next chapter begins to introduce the proposed deterministic localisation method, Intersection after Principal Component Analysis (PCA-Intersection), to estimate user's location based on the proposed clustering scheme.

Chapter 5

Deterministic Estimation with Clustering

5.1 Introduction

As described previously in chapter 4, in the training phase, after the number of clusters is identified, PCA approach is used to rotate the raw RSS to independent principal components in each determined cluster. Since the RSS values in each cluster are similar, the application of PCA can retain accuracy by not losing the substantial RSS correlations in each cluster, but also the PCA accommodates the different RSS distributions in each cluster. This allows building RSS distribution models within each cluster to support further positioning. This chapter concentrates on how to model the transformed RSS distributions in each cluster and how to estimate a user's location by finding the most likely intersection area of more than three BSs circles in geographical spaces. The detail of the *Intersection after Principal Component Analysis* (PCA-Intersection) method is described in section 5.2. Then in section 5.3, the proposed method is compared with the KNN method using three forms of partitioning, viz. the whole area (a.k.a global partitioning or no partitioning), cluster partitioning and grid partitioning to evaluate the

positioning performance based on the simulated and real data sets discussed in chapter 4. Additionally, the reasons for using deviation RSS data and the Mahalanobis distance function are analysed in section 5.4 and section 5.5 respectively. Section 5.6 concludes this chapter.

5.2 Intersection after Principal Component Analysis Method

5.2.1 Training Phase

In chapter 4 section 4.5, the transformed training data $R_i^{'}$ in each cluster C_i can be obtained. After getting the $R_i^{'}$, the propagation model in each cluster C_i is built in order to estimate the distance of a new MS. Let suppose there are n_i MSs in the cluster C_i. Here the transformed RSS data is used to describe the "distance" between the BS b and the MS j (the j-th transformed training data in $R_i^{'}$ with the following function:

$$d(r_{j,b}^{'}) = 10^{(PTR - r_{j,b}^{'})/10\alpha_b} \tag{5.1}$$

Where $r_{j,b}^{'}$ is the transformed signal power from BS b to MS j and PTR is the value of the transmission power and in the experiments, which is given a default value of 48 dBm for an outdoor GSM environment. The parameter α_b for BS b can be chosen to minimize the sum of the squared errors $\sum_{j=1}^{n_i}(d_{j,b} - d(r_{j,b}^{'}))^2$ where $d_{j,b}$ is the distance calculated from the already known locations of BS b and MS j from the training data.

5.2.2 Online Location Estimation Phase

In this phase, the location of a new MS is estimated. Suppose that for a new MS m the observed RSS tuple from q neighbouring BSs is $\vec{r}_m = (r_{m,1}, r_{m,2}, ..., r_{m,q})$. The detailed process of its location estimation is shown below:

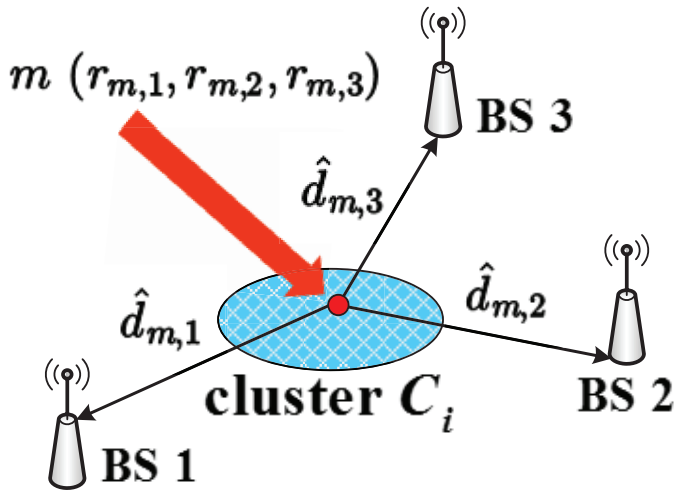


Figure 5.1: RSS distribution models built for one cluster

Step 1: Use the VPM based on KNN to calculate the probability of MS m belonging to each cluster, and then assign the cluster ID with the highest probability to MS m. Call this cluster C_i.

Step 2: Transform the observed RSS tuple of MS m into the new basis of cluster C_i with its PC coefficients A_i. Thus getting $\vec{r}'_m = A_i^T \cdot (\vec{r}_m - \bar{R}_i)$, and $\vec{r}'_m$, and $\vec{r}'_m$ can also be represented as $(\vec{r}'_{m,1}, \vec{r}'_{m,2}, ..., \vec{r}'_{m,q})$.

Step 3: Based on the function (5.1) and the optimized parameter α_b for each BS, the "distance" between each BS to MS m can be estimated as $\hat{d}_{m,b} (1 \leq b \leq q)$, as shown in Figure 5.1. (Here $q = 3$).

Step 4: Calculate the distance in signal space between $\vec{r}'_m$ and all the transformed training data set in cluster C_i. Then KNN algorithm is used to find MS m's K nearest neighbours. These neighbours are listed as $\{p_1, ..., p_k, ..., p_K\}$, which is arranged in descending order of signal distance. In order to estimate of the precision for each calculated distance, a distance range is computed with respect to each BS, viz. $\delta_{m,b}$ for BS b. It is important to calculate the confidence level of the distance band. To do this, MS m's K neighbours' individual deviations from the distances from the centroid of the K

neighbours in this cluster are calculated, this is given by:

$$\delta_{k,b} = \sqrt{|d_{p_k,b} - \overline{d_{p,b}}|} \tag{5.2}$$

Here $d_{p_{k,b}}$ is the distance calculated from the already known locations of BS b and MS p_k. $\overline{d_{p,b}}$ is the centroid of the K nearest neighbours in training data points from cluster C_i and it is defined as:

$$\overline{d_{p,b}} = \sum_{k=1}^{K} d_{p_k,b}/K \tag{5.3}$$

Then the weighted mean and standard deviation of the deviations $\delta_{p,b}$ are calculated. Let μ be the mean value of the deviations, and σ the standard deviation of the deviations. Since the RSS distribution is skew in the real environment, the confidence interval derived from the normal distribution cannot be used in this case. Therefore, a two sided confidence interval can be created by applying Chebyshev's inequality [89], which can provide a lower bound for how much probability mass lies outside a the chosen confidence range. As such, the two-sided confidence level of the distance band interval can be obtained by:

$$P(|\delta_{p,b} - \mu \geq \lambda\sigma|) \leq \frac{1}{\lambda^2} \tag{5.4}$$

On the right hand side of (5.4), the value of $\frac{1}{\lambda^2}$ is set to be 0.01, which means that whatever the distribution is, there is always at least 99% of the probability of being inside the distance band interval. The value of the upper bound of the band width as the uncertainly band, $\delta_{m,b}$ is now chosen. Thus, the possible distance $d_{m,b}$ between MS

m and BS b is given by:

$$d_{m,b} \in \left[\hat{d}_{m,b} - \delta_{m,b}, \hat{d}_{m,b} + \delta_{m,b}\right] \tag{5.5}$$

Step 5: At this step, the aim is to find out which one of the intersection areas MS m is most likely to be located in. Different intersection areas are generated by different patterns of distance bands of at least two BSs. This is illustrated in Figure 5.2 that, for simplicity, only shows three BSs. A similar idea to [90] is adopted for the search. Since the distribution of training data points is quite variable, the search strategy considers the two cases below.

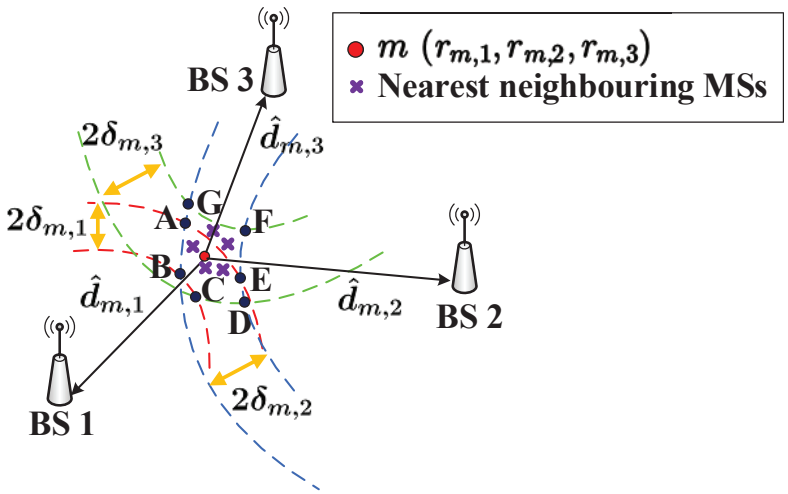


Figure 5.2: Uncertainty area of location estimation

1. The intersection area that has the most number of nearest neighbours is selected. For example, there are 5 nearest neighbours of MS m in Figure 5.2. The intersection area ABCDEA has three neighbours and the area AEFGA has two, so the most likely intersection area for MS m is taken to be ABCDEA.

2. If more than one intersection area has the same greatest number of nearest neighbours (including the case where none of the intersection areas contains even one of these neighbours), choose the area that contains the neighbour points with the

smallest sum of distance in signal strength between the neighbour's RSS values and the MS m's RSS values.

Step 6: Collect all the training data points in cluster C_i that are located on the most possible intersection area, and use the WKNN method to compute MS m's location. Assume there are $K^{'}$ qualified training data points, and $(\vec{r}'_1, ..., \vec{r}'_i, ..., \vec{r}'_{K'})$ and $(l_1, ..., l_i, ..., l_{K'})$ denote their RSS sets and locations respectively. The location of MS m is estimated by $\hat{l}_m = \sum_{i=1}^{K'} w_i l_i$. The w_i is a normalized weight for each training data point and is given as: $w_i = \frac{1}{\|\vec{r}'_i - \vec{r}'_m\| \cdot \sum_{i=1}^{K'} \frac{1}{\|\vec{r}'_i - \vec{r}'_m\|}}$. Figure 5.3 provides a sample to display how the method above process with a real training data set.

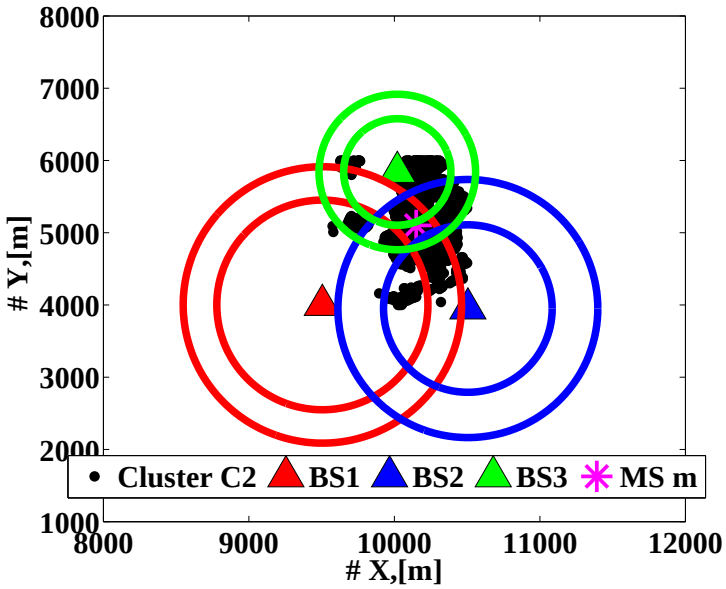


Figure 5.3: A sample of the uncertainty area

5.3 Performance Evaluation

In this section, the localisation accuracy is tested with four different sets of data. In all cases, the MS were stationary or walking. Three forms of partitioning are considered, viz. the whole area (a.k.a global partitioning or no partitioning), cluster partitioning

and grid partitioning. For the sake of comparison, the number of grid elements is made to correspond to the number of clusters generated. To better compare the performance between the proposed method and the traditional KNN method, in the experiments for each test-bed, the data points are randomly divided into two equal sets. The first half is treated as training data points and their location coordinates are assumed to be known. The other half is used for location estimation with only the RSS values, with their location information (e.g. GPS values) subsequently used for validation. The collected GPS data is assumed accurate in this thesis, so it is used as a reference. All the estimated locations using different algorithms are compared with the reference GPS data to calculate the root mean square errors (RMSE). Therefore, the algorithm which gave the estimation with the minimal RMSE is considered as "good".

5.3.1 Location Results with Simulated Data Sets

5.3.1.1 Outdoor Scenario 1: A Simple Simulated Urban Propagation Model

Figure 5.4 depicts the cumulative distribution function (CDF) of the error distance of the KNN and PCA-Intersection algorithms based on global, grid and cluster models. Although the propagation model in outdoor scenario 1 is established based on grid elements, the results show that the cluster-based positioning methods provide a slightly better accuracy than the grid-based positioning methods and no partitioning-based positioning methods. In addition, Table 5.1 summarizes the information in terms of the mean, 50th, 75th and 90th percentile values of the error distance for these two algorithms based on the three models. For example, 90th percentile of the distance errors using cluster-based PCA-Intersection method is within 219.6 m whereas the method based on grid model and global model report 299.9 m and 440.2 m respectively. For PCA-Intersection algorithm, cluster-based leads to improvements of 23.4 m over using grid model. It indicates that using clustering scheme can give good support for the appropriate estimation of the mobile users' locations even with a grid-based propagation model.

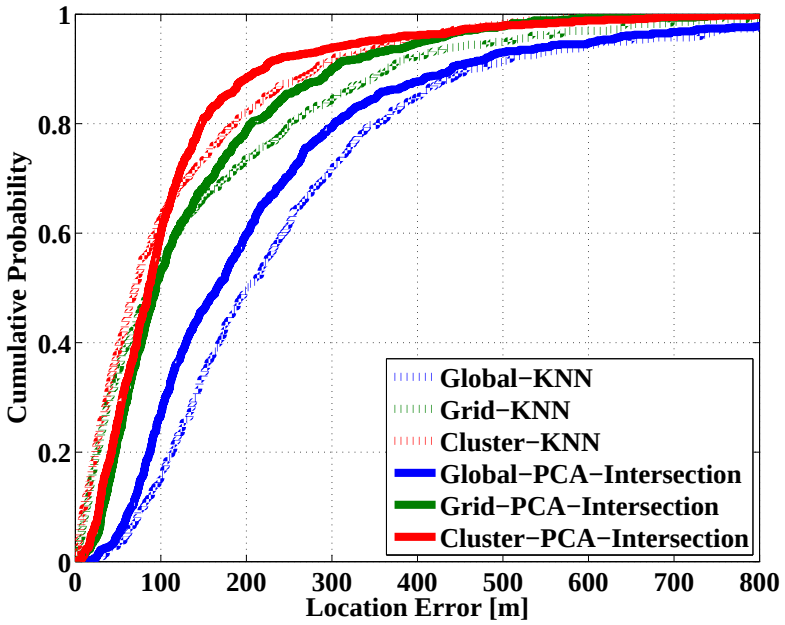


Figure 5.4: Cumulative percentile of error for KNN and PCA-Intersection algorithms based on different partitioning models in outdoor scenario 1: a simple simulated urban propagation model.

Table 5.1: Comparison of Estimation Error between KNN and PCA-Intersection Methods based on Global, Grid and Cluster Models in Outdoor Scenario 1 (in meters)

	Outdoor Scenario 1: A Simple Simulated Urban Propagation Model					
	KNN			PCA-Intersection		
	Global	Grid	Cluster	Global	Grid	Cluster
Mean Error	252.2	148.3	114.7	216.4	137.4	114.0
50 Percentile	201.2	86.2	70.1	166.0	94.1	85.9
75 Percentile	316.8	211.9	156.1	267.6	182.0	133.8
90 Percentile	463.1	364.7	282.0	440.2	299.9	219.6

5.3.1.2 Outdoor Scenario 2: The Island of Jersey

As shown in Figure 5.5, the proposed algorithm also outperforms the KNN algorithm by using a cluster model with a suitable number of clusters. Similarly, as illustrated in Table 5.2, it can be observed that for the proposed algorithm (Cluster-PCA-Intersection), 90th percentile of the distance errors is within 26.8 m, whereas Cluster-KNN method

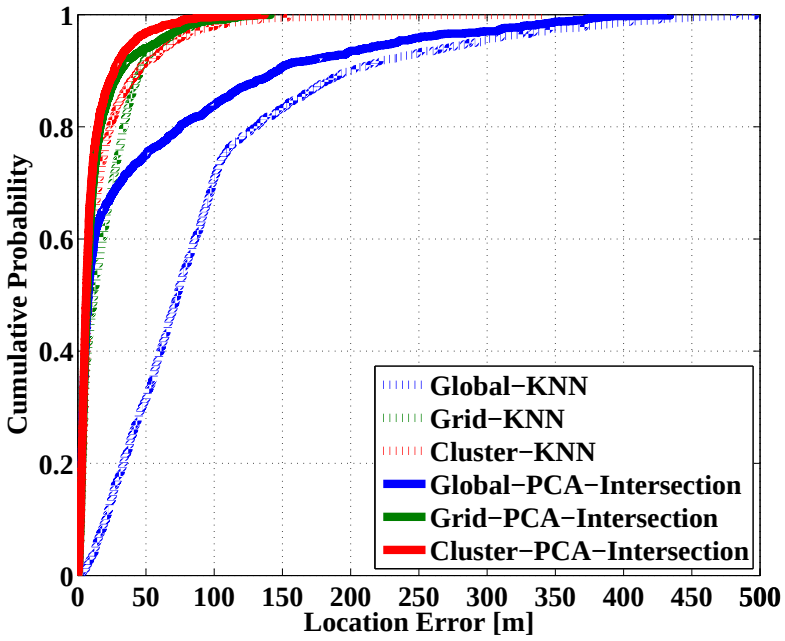


Figure 5.5: Cumulative percentile of error for KNN and PCA-Intersection algorithms based on different partitioning models in outdoor scenario 2: the island of Jersey area.

Table 5.2: Comparison of Estimation Error between KNN and PCA-Intersection Methods based on Global, Grid and Cluster Models in Outdoor Scenario 2 (in meters)

	Outdoor Scenario 2: The Island of Jersey area					
	KNN			PCA-Intersection		
	Global	Grid	Cluster	Global	Grid	Cluster
Mean Error	95.9	19.6	13.4	45.5	17.6	11.2
50 Percentile	76.5	12.0	6.75	7.8	8.0	6.1
75 Percentile	108.0	29.5	13.7	49.7	19.2	11.8
90 Percentile	204.4	45.2	31.2	146.3	44.6	26.8

report 31.2 m to reach the same cumulative probability. Likewise, in the grid model, the PCA-Intersection algorithm achieves a little better positioning accuracy than the KNN algorithm. 90th percentile of the positioning errors for Grid-PCA-Intersection is with 44.6 m, which is similar to (slightly better than) the result of Grid-KNN algorithm in the rural environment. It can be seen from the results from this scenario, the positioning accuracy performances of the two algorithms in terms of cluster model are quite similar

to those in terms of grid model in the rural propagation environment.

In summary, given the accuracy of location estimation and the complexity of the model it seems feasible to adopt the algorithms proposed in this chapter to estimate location based on the clustering scheme that can have a mapping to the topography meaningful in a large area.

5.3.2 Location Results with Real Data Sets

5.3.2.1 Outdoor Scenario 3: Queen Mary campus

Figure 5.6 and Table 5.3 show the CDF of the error distance and the position error using the PCA-Intersection and KNN based on global model, grid model and cluster model respectively. The PCA-Intersection and KNN based on the cluster model outperforms these two methods based on the other two models. In particular, for the cluster-based PCA-Intersection approach, 50th percentile of the distance errors is within 29.4 m, and the distance measurement error is around 75.6 m in mean value.

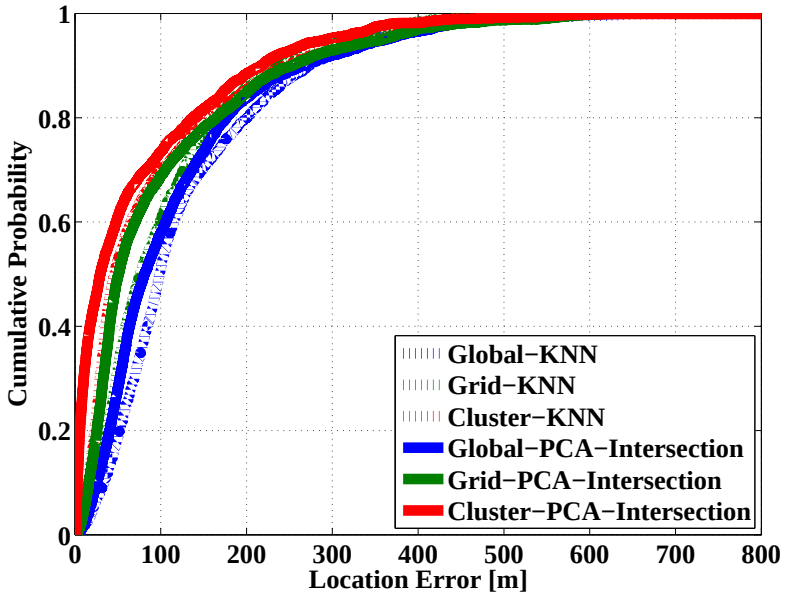


Figure 5.6: Cumulative percentile of error for KNN and PCA-Intersection based on different partitioning models in outdoor scenario 3: Queen Mary campus.

Table 5.3: Comparison of Estimation Error between KNN and PCA-Intersection Methods based on Global, Grid and Cluster Models in Outdoor Scenario 3 (in meters)

	Outdoor Scenario 3: Queen Mary Campus					
	KNN			PCA-Intersection		
	Global	Grid	Cluster	Global	Grid	Cluster
Mean Error	131.4	107.9	95.0	119.1	99.2	75.6
50 Percentile	98.4	76.3	46.0	83.1	49.5	29.4
75 Percentile	172.4	142.9	123.6	156.4	130.6	106.3
90 Percentile	275.6	237.6	244.1	262.4	251.9	216.9

Unlike the rural environment in outdoor scenario 2, the signal in complex suburban environment undergoes additional attenuation and fluctuates rapidly due to many obstructions i.e. high buildings. The clustering scheme uses deviation signal strengths to partition the environment into consistent geographic regions which are more homogenously covered by the radio signal. The intention is to better model a realistic complex environment than with a grid model. These results not only indicate that the cluster based positioning methods outperform grid based methods in the complex outside area, but also the proposed intersection method can provide relatively high location estimation accuracy with only a small amount of training data points in a small area. Since the ground truth is taken as the GPS data and it has inaccuracies, the variance is an overestimate.

5.3.2.2 Outdoor Scenario 4: Music Festival in London Victoria Park

In the Day 1 data set, 1048 test points were randomly selected as training data points and the remaining test points used as testing data set. The same size of training data set is used in each method. Similarly, as illustrate in Figure 5.7, the results again indicate that the proposed method (using clustering) provides greater positioning accuracy in the suburban environment. In Table 5.4, it can be observed that for the Cluster-PCA-Intersection algorithm, 50th percentile of the distance errors is within 48.6 m, whereas

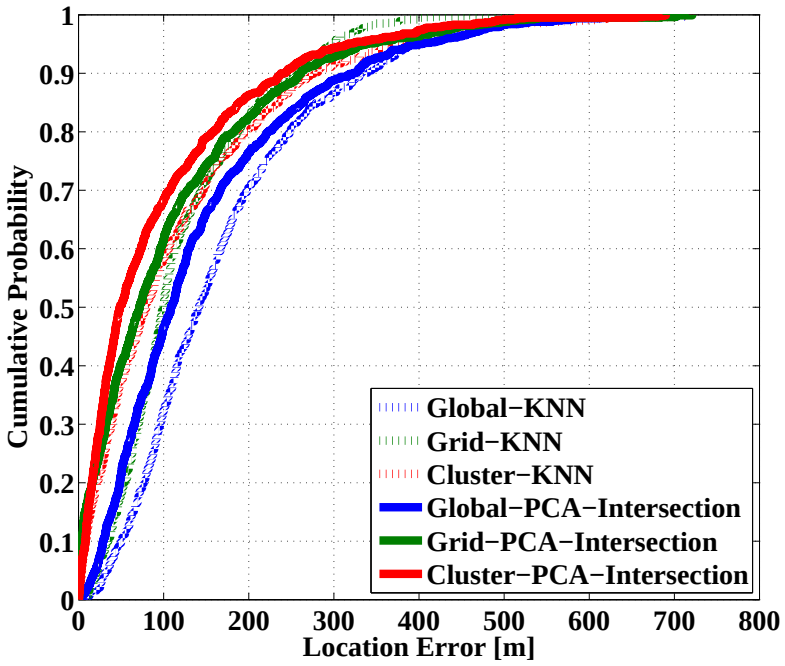


Figure 5.7: Cumulative percentile of error for different algorithms based on different partitioning models in outdoor scenario 4: Music Festival in London Victoria Park on Day 1.

Table 5.4: Comparison of Estimation Error between KNN and PCA-Intersection Methods based on Global, Grid and Cluster Models in Outdoor Scenario 4 (in meters)

	Outdoor Scenario 4: Music Festival in London Victoria Park					
	KNN			PCA-Intersection		
	Global	Grid	Cluster	Global	Grid	Cluster
Mean Error	168.7	122.1	119.0	143.7	108.2	93.6
50 Percentile	141.9	99.4	79.6	109.0	71.7	48.6
75 Percentile	221.7	165.6	164.2	193.1	154.0	131.4
90 Percentile	337.9	248.6	278.7	323.2	261.0	240.3

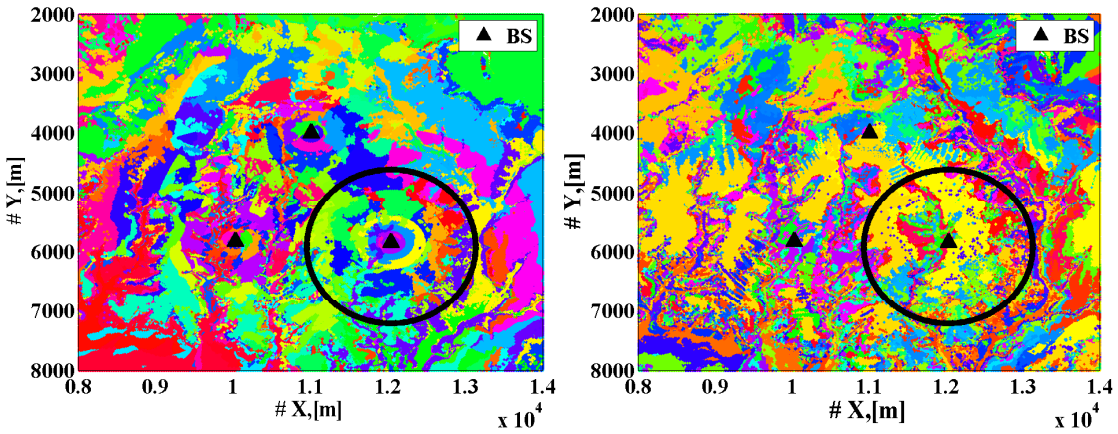
the Cluster-KNN method reports 79.6 m to reach the same cumulative error probability. Likewise, for the grid partitioning, the intersection method also achieves a higher positioning accuracy than the KNN method. The mean value of positioning error for Grid-PCA-Intersection is within 108.2 m, which is a little broader than the result for the Cluster-PCA-Intersection method in the suburban environment. From the results in

this scenario, it can be seen that the positioning accuracy performance of the two algorithms for cluster partitioning follow the same pattern to the grid model in the suburban propagation environment.

5.4 RSS deviations from path loss versus raw RSS

Using the deviations helps in two ways:

(a) Clusters generated by the deviations data are a better reflection of the topography. For the island of Jersey data, Figure 5.8 depicts the comparisons of clustering distribution between using the raw RSS and deviation RSS when the same number of clusters is created. It can be seen that using the deviation RSS has achieved significant better result than raw RSS. Using the raw RSS leads to clusters where the similarity is dominated by the distance path loss when near a BS and this is approximated anyway by the estimated path loss model. The evidence for this is observed in Figure 5.8 (a) (especially for the black circle area) when the raw RSS was used. For example, if the world were uniform, then ring segments are generated. This would simply reflect the decay with distance rather than topography. On the other hand, Figure 5.8 (b) illustrates that when using



(a) Clustering based on raw RSS

(b) Clustering based on deviation RSS

Figure 5.8: The comparisons of clustering results between using the raw RSS and deviation RSS in the island of Jersey data

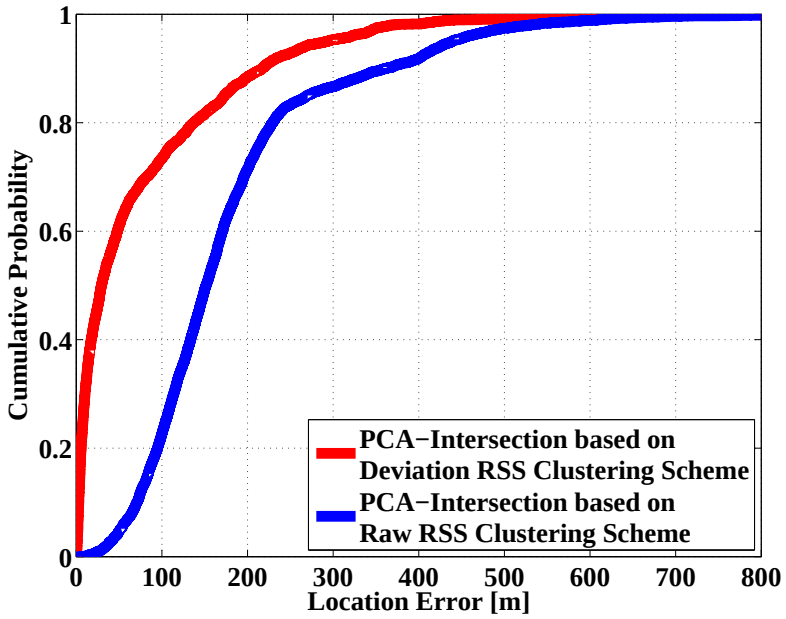


Figure 5.9: Location estimation results based on raw RSS and deviation RSS clustering scheme in Queen Mary data by using PCA-Intersection approach

the deviations the mapping of the clusters onto the geographical locations shows more scatter and reflects the terrain better than if the clustering was performed on the raw data in Figure 5.8 (a).

(b) The location estimation is more accurate on all the data sets when using the deviations, e.g. Figure 5.9 shows an example of the CDF of the error distance for Intersection method based on both data sets in the Queen Mary Scenario, under the premise that the same number of clusters is created. For the Intersection method the mean values of the distance error based on deviations is 75.6 m, whereas based on the raw RSS data the mean is 181.0 m. The invariance of the clusters to transmit powers using the deviations was illustrated in outdoor scenario 1 in chapter 4 section 4.6.1.1.

5.5 Mahalanobis's Distance versus Euclidean Distance

The correlation between signal strength is very common in a real environment, which has an impact on the estimation accuracy. For clustering, using the Mahalanobis distance

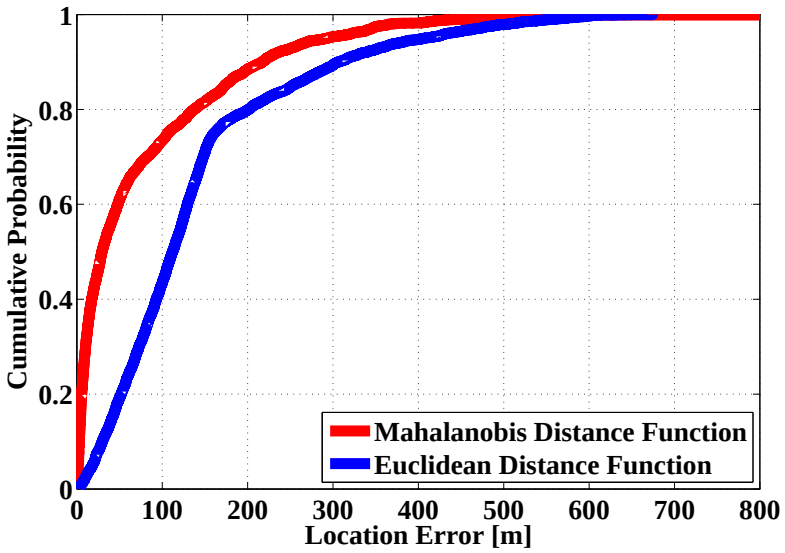


Figure 5.10: Cumulative percentile of error for cluster-based approach using Mahalanobis distance and Euclidean distance for the Queen Mary campus data

function to calculate the RSS similarity between any two MSs from different transmitters can correct for the high correlation between signal strength from different transmitters and automatically account for the scaling of the coordinate axes. Figure 5.10 compares the CDF of the error distance for the proposed probabilistic method based on both data sets in the Queen Mary Scenario, under the premise that the same number of clusters is created. It can be observed that the proposed method based on clustering using Mahalanobis distance function significantly outperforms that based on clustering using Euclidean distance. More specifically, for the proposed method the mean values of the distance error based on Mahalanobis distance function is 75.6 m, whereas based on the Euclidean distance function the mean is 140.8 m. This result indicates that the Mahalanobis distance could be used as a better alternative to the Euclidean distance to determine the location in positioning systems. The improvement provided by the Mahalanobis distance lies in its use of the covariance matrix to adjust the distance metric. The Euclidean distance essentially treats all the data points equally. Two highly correlated RSS values, for example are given too much weight. The Mahalanobis function corrects for this.

5.6 Summary

An improved intersection localisation method to outdoor fingerprint location estimation based on clustering RSS from BSs has been presented in this chapter. Four different scenarios (rural, urban, and suburban) have been considered in order to evaluate the performance of the proposed approach. Results presented show the proposed scheme finds more accurate locations and outperforms the KNN approaches for all numbers of NNs tested based on global model, cluster model and grid model in the four test-beds. It is also shown that the difference between clustering and the more conventional grid partitioning is important in urban environments, as in these more complex environments the regular partitioning of grids leads to higher variability in the accuracies obtained. Moreover, the experimental results also showed that the clustering scheme created by using deviations in RSS from the estimated path loss gives a better reflection of the topography and this clustering based on Mahalanobis distance is seen to provide better accuracy than that based on the Euclidean distance in a complex environment.

The following chapter will focus on calculating mobile users' locations by using a probabilistic algorithm based on the proposed clustering scheme.

Chapter 6

Probabilistic Estimation with Clustering

6.1 Introduction

In this chapter, the main focus is on the use of probabilistic models for location estimation. Section 6.2 introduces an overview of the probability framework and points out the estimation problem by using probability techniques in previous works. The general concept of the Kernel Density Estimation (KDE) technique is described in section 6.3. Section 6.4 proposes a novel estimator derived through KDE technique by using Principle Component Analysis (PCA) to transform highly correlated RSS values into uncorrelated adjusted RSS values. In section 6.5, the proposed PCA-KDE technique is compared with the common existing KDE method and performance are evaluated based on the simulated and real data sets discussed in chapter 4. Finally, section 6.5 concludes of this chapter.

6.2 Probabilistic Framework

The main idea of the probabilistic framework is to calculate the conditional probability density function (pdf) $P(l_i|r)$, $i = 1, ..., l$ (posterior) given the observed RSS fingerprint r during positioning, which can be done by using Bayes' Theorem:

$$P(l_i|r) = \frac{P(r|l_i)P(l_i)}{P(r)} = \frac{P(r|l_i)P(l_i)}{\sum_{l_i \in L} P(r|l_i)P(l_i)} \tag{6.1}$$

Where the function $P(r|l_i)$ is the likelihood function of the given RSS measurement $r = r$, $P(l_i)$ is the prior probability of being at location l_i before knowing the value of the observed RSS and $P(r)$ is the normalizing constant. Usually, the prior density $P(l_i)$ is assumed to be uniform distribution, thus the problem is to calculate $P(r|l_i)$, though knowledge of hotspot clusters in an area would allow non uniform priors.

In previous indoor literatures, most of them assume the RSS measurements from APs are independent. If there are n transmitters, e.g. APs or BSs, the value of $P(r|l_i)$ can be computed as:

$$P(r|l_i) = \prod_{\rho=1}^{n} P(r_\rho|l_i) \tag{6.2}$$

Here n is the number of transmitters that are used to form the RSS tuple. The probability $P(r_\rho|l_i)$ can be obtained by calculating the pdf of the RSS measurements from the ρ-th transmitter at each of the training locations l_i. This can be estimated from the training data set in the training phase. The information about the PDFs of the RSS training data points are retained in approximating functions, e.g. in in the form of histograms or as a kernel function. In this thesis, the kernel functions are considered and more detail is given in the next section. Finally, the use of the Maximum Likelihood (ML) estimator, Maximum A Posteriori (MAP) estimator and Minimum Mean Square Error (MMSE) estimators is to calculate the location of a desired MS, as illustrated in Table 6.1, are described. If the prior distribution is uniform, the MAP estimate is the

same as the ML estimate.

Table 6.1: Different Positioning Variants

Name	Function
Maximum Likelihood (ML)	$\hat{l} = \text{agrmax}_{l_i} P(r\|l_i)$
Maximum A Posteriori (MAP)	$\hat{l} = \text{agrmax}_{l_i} P(r\|l_i)P(l_i)$
Minimum Mean Square Error (MMSE)	$\hat{l} = E(l\|r) = \sum_{i=1}^{L} l_i P(l_i\|r)$

6.3 Introduction of Kernel Method

The kernel density estimator (KDE) can be also called a Parzen Window estimator [91]. It is an alternative to the histogram for nonparametric density estimation. Compared to the histogram, the KDE uses a smooth and continuous function for building blocks for bins, and also enjoys superior theoretical properties such as integrated variance [3]. Intuitively, the KDE is the superposition of "bumps" centred at each training data point [91]. The shapes of these "bumps" are determined by a kernel function $K(\cdot)$, depicted in Figure 6.1. The kernel function is a non-negative function satisfying the following conditions:

$$\int K(r)dr = 1 \tag{6.3}$$

$$K(r) = K(-r) \tag{6.4}$$

Hence, the kernel estimator with kernel $K(\cdot)$ is defined by [91]

$$f(r) = \frac{1}{n\delta} \sum_{i=1}^{n} K(\frac{r - r_i}{\delta}) \tag{6.5}$$

Here n is the number of data points. δ is the window width, also called the smoothing parameter or bandwidth. The window width greatly affects the resulting density

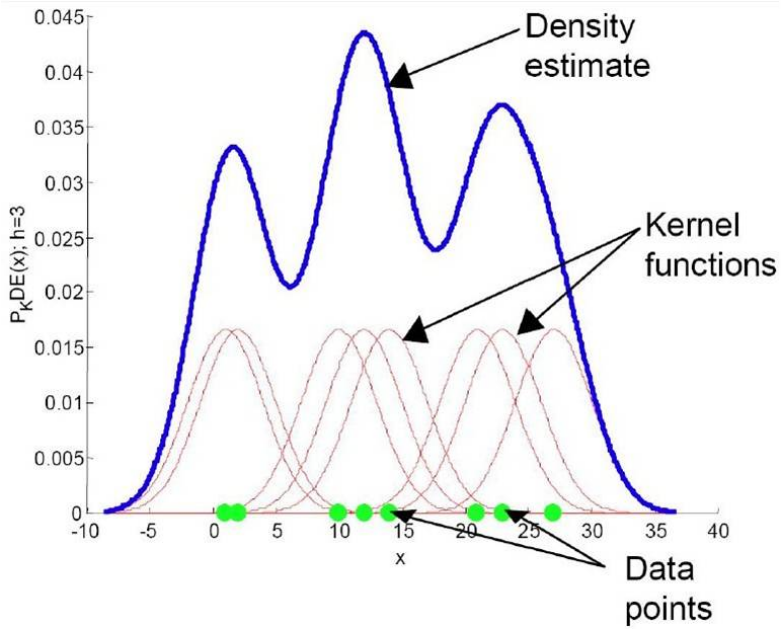


Figure 6.1: Kernel density estimate as a sum of bumps [5]

estimate. The quality of a density estimate is now widely recognized to be primarily determined by the choice of the smoothing parameter δ, and only in a minor way by the choice of kernel function.

6.3.1 Kernel Function

Different kernel functions generate different shapes of the estimated density. Table 6.2 lists four examples of kernel functions. Since the Gaussian kernel performs effectively in previous indoor positioning applications, it is also used here. In [3], the authors compare the influence of the kernel function in Table 6.2 in the final estimates, and suggest that the shape of the kernel does not lead to drastic changes in the accuracy of the estimate.

[5] This figure is made by Ricardo Gutierrez-Osuna, Wright State University, and the parameter h is the window width.

Table 6.2: Kernel Functions [3]

Kernel	Functions
Triangle	$K(r) = 1 - \|r\|, \|r\| \leq 1$
Epanechnikov	$K(r) = \frac{3}{4}(1 - r^2), \|r\| \leq 1$
Quartic	$K(r) = \frac{15}{16}(1 - r^2)^2, \|r\| \leq 1$
Gaussian	$K(r) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}r^2)$

6.3.2 Kernel Bandwidth

Choosing a good bandwidth is crucial in density estimation. Choose too large bandwidth, the kernel function cannot reflect the exact density distribution as some important information might be neglected as the spread smoothes over important features of the overall density function. If the bandwidth is too small, meaningful information may suffer significant statistical fluctuation because of the paucity of samples in each bandwidth. Figure 6.2 presents examples of Kernel densities with various bandwidths chosen, and illustrates how increasing of the bandwidth can simplify the Kernel densities and level out minor variations.

In fact, a number of alternative measures exist to estimate bandwidth δ and can be found in [3]. The appropriate choice for the value of δ is data-dependent and also depends on how the density estimates are to be used. In this thesis, the optimal δ is obtained by minimizing the asymptotic mean integrated square error (AMISE) between the estimated and true densities [3], which is given by

$$\delta = \{\frac{4}{n(2d+1)}\}^{1/(d+4)}\sigma \tag{6.6}$$

Here d is the dimension of the RSS vector and $\sigma^2 = \frac{1}{d}\sum_{k=1}^{d}\sigma_k^2$ is the average marginal

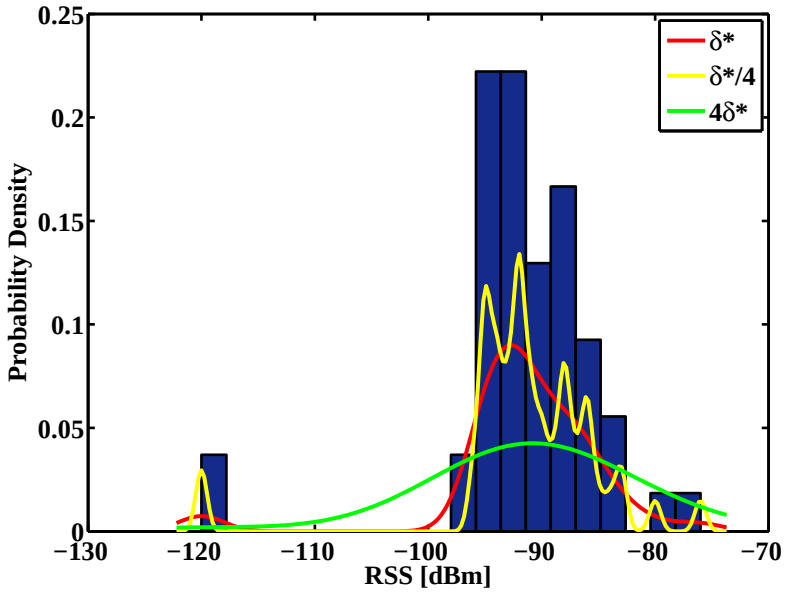


Figure 6.2: Kernel density estimator with different bandwidths using Queen Mary data set

variance in each dimension.

6.4 Constructing Kernel Density Estimator after Principle Component Analysis

6.4.1 Training Phase

After the RSS transformation process as described in chapter 4 section 4.5, the probability density estimate is calculated for each of the independent axes in the transformed training RSS samples $\vec{t}_b^{'}$, $\vec{t}_b^{'} \in R_i^{'}$ from the BS b in location l ($l \in L_i$). Here the Gaussian KDE method is utilized as shown in (6.7):

$$P(r_{j,b}^{'}|l = l_{i,j}) = \frac{1}{n_i} \sum_{u=1}^{n_i} \frac{1}{\sqrt{2\pi}\Psi_b} \exp\left(- \frac{(r_{j,b}^{'} - r_{u,b}^{'})^2}{2\Psi_b^2} \right) \tag{6.7}$$

Ψ_b is the smoothing parameter that determines the width of the kernel. The value of

Ψ_b is not only data-dependent but also depends on how the density estimates are used. Here the optimal Ψ_b is obtained according to (6.6).

6.4.2 Online Localisation Phase

At run time: Given a new MS m with observed RSS tuple $\vec{r}_m$ from q BSs, the process of estimation of MS m's location ($\hat{l}_m$) is as follows:

Step 1 and 2: These are the same steps as the Intersection method. Predict the estimated cluster ID for the new MS m and then transform its RSS tuple into PCs from (4.16), Here $\vec{r}'_m = A_i^T \cdot (\vec{r}_m - \bar{R}_i)$ and $\vec{r}'_m = (r'_{m,1}, r'_{m,2}, ..., r'_{m,q})$.

Step 3: Estimate the probability of $\vec{r}'_m$ over all possible training location values in cluster C_i. The posterior probability density function of the location l is given by Bayes' Theorem:

$$P(l_{ij} \,|\, r = \vec{r}'_m) = \frac{P(r = \vec{r}'_m \,|\, l_{ij}) P(l_{ij})}{\sum_{l_{ij} \in L_i} P(r = \vec{r}'_m \,|\, l_{ij}) P(l_{ij})} \tag{6.8}$$

Since the only aim is to select the most probable location rather than compute the actual probability, the denominator can be ignored as it is the same for all the possible locations l_{ij} ($l_{ij} \in L_i$) in the training samples within cluster C_i. Here other factors, such as relative traffic densities, are not taken into consideration (though they could be) and the prior density $P(l_{ij})$ is regarded to be uniform in the cluster.

Estimating $P(r = \vec{r}'_m | l_{ij})$ requires calculating the frequency of $\vec{r}'_m$ for every possible location l. In the training step, the kernel probability distributions for each dimension for each training location in each cluster has been built. Therefore, the probability of $\vec{r}'_m$ at every possible location l_{ij} can be obtained, as the product of the conditional probability density function (from the kernel function) of each dimension:

$$P(r = \vec{r}'_m \,|\, l_{ij}) = \prod_{b=1}^{q} P(r_b = r'_{m,b} \mid l_{ij}) \tag{6.9}$$

This also can be considered as a weight of $\vec{r}'_m$ at the location l_{ij}. For the subsequent location estimation of MS m it is important to normalize all the weights. Let $w_{m,l_{ij}}$ be the normalized weight of MS m at the location l_{ij}, so $w_{m,l_{ij}} = \frac{P(r=\vec{r}'_m|l_{ij})}{\sum_{j=1}^{n_i} P(r=\vec{r}'_m|l_{ij})}$. Thus, the estimated location of MS m is $\hat{l}_m = \sum_{j=1}^{n_i} w_{m,l_{ij}} l_{ij}$.

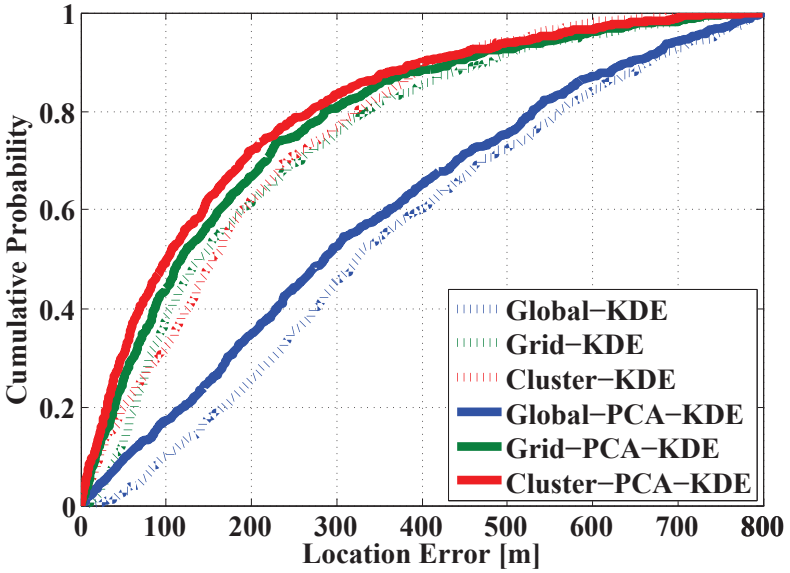
6.5 Experimental Results

6.5.1 The Comparisons of Different Partitioning Models

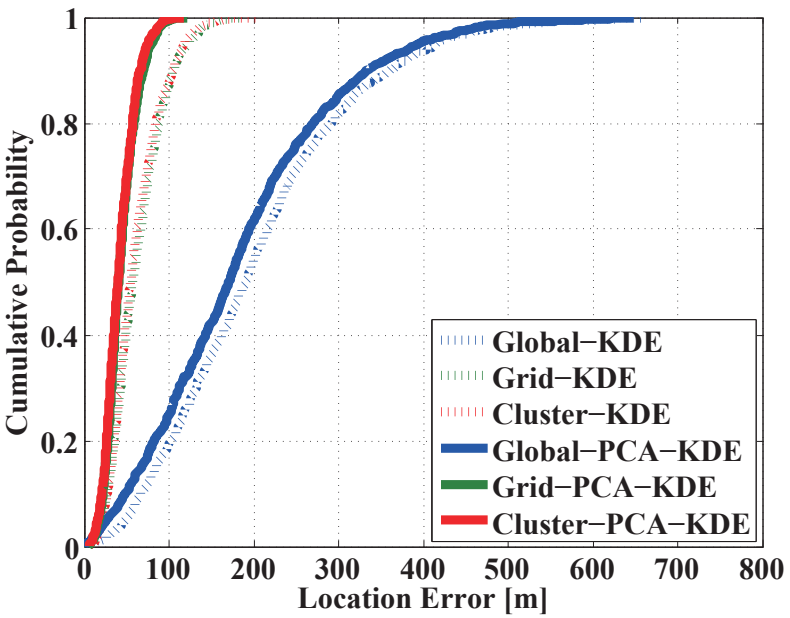
In this section, the proposed method is compared with the traditional KDE method by using the data set from the simulated and real environments discussed in chapter 4. Similarly, every method is tested by using different partitioning models (global-, grid- and cluster-) and the number of generated grid elements is equal to the number of cluster created. In the experiments for each test-bed, the data points are randomly divided into two equal sets. The first half is treated as the training data set and their location coordinates are assumed known. The other half is used for location estimation with only the RSS values, with their location information (e.g. GPS values) subsequently used for validation.

6.5.1.1 Location Results with Simulated Data Sets

Using the simulated data set in outdoor scenario 1 and outdoor scenario 2 in chapter 4, the proposed method is compared against the original KDE approaches based on global, cluster and grid partitioning models. The positioning error information between these two methods based on different models are illustrated in Table 6.3 and the CDFs of the error distance of them are shown in Figure 6.3. As previously described, the propagation model of outdoor scenario 1 is built based on uniform grid elements. It seems reasonable that approaches based on a grid model would provide similar or slightly better positioning results than these methods based on a cluster model. The simulation results



(a) Outdoor Scenario 1: A Simple Simulated Urban Propagation Model



(b) Outdoor Scenario 2: The Island of Jersey area

Figure 6.3: Cumulative percentile of error for KDE and PCA-KDE based on different partitioning models in simulated environments: (a) A Simple Simulated Urban Propagation Model; (b) The Island of Jersey area

Table 6.3: Comparison of Estimation Error between KDE and PCA-KDE Methods based on Global, Grid and Cluster Models in Outdoor Scenario 1 and 2 (in meters)

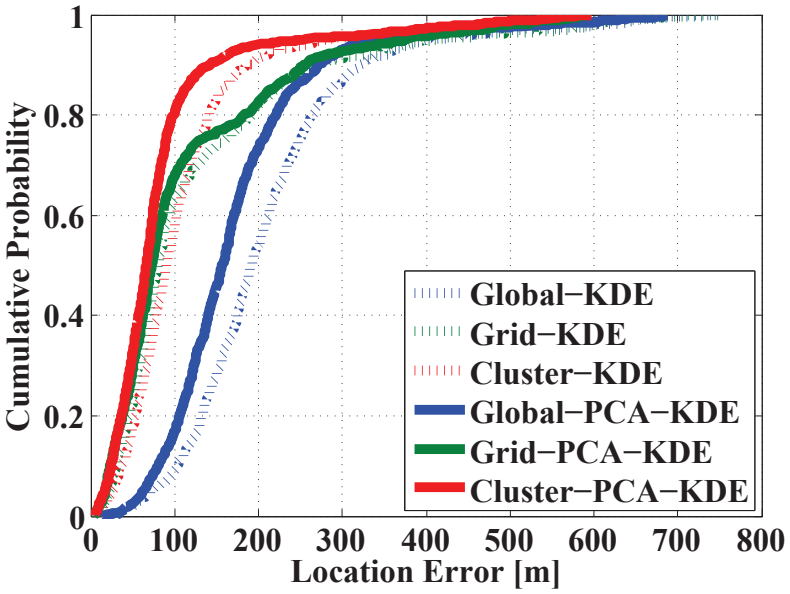
	Outdoor Scenario 1: A Simple Simulated Urban Propagation Model					
	KDE			PCA-KDE		
	Global	Grid	Cluster	Global	Grid	Cluster
Mean Error	360.8	204.2	194.1	323.2	177.4	159.8
50 Percentile	323.3	137.6	154.9	286.3	117.9	100.7
75 Percentile	526.6	296.8	286.6	477.7	252.8	225.5
90 Percentile	668.0	447.7	416.3	644.3	436.0	400.5
	Outdoor Scenario 2: The Island of Jersey area					
	KDE			PCA-KDE		
	Global	Grid	Cluster	Global	Grid	Cluster
Mean Error	202.6	62.7	60.7	183.2	43.4	41.8
50 Percentile	191.6	57.4	55.0	170.4	40.7	39.5
75 Percentile	265.8	81.7	79.0	246.9	54.3	53.6
90 Percentile	361.1	105.4	104.3	333.0	70.0	65.7

show that the PCA-KDE method outperforms the original KDE method to calculate estimated location information, no matter which one of partitioning models it is applied to. In terms of mean square error, Cluster-based PCA-KDE, Grid-based PCA-KDE and Global-based PCA-KDE methods report values of the distance errors that are within 159.8 m, 177.4 m and 323.2 m respectively, whereas the original KDE method based on clustering scheme, grid model and global model are within 194.1 m, 204.2 m and 360.8 m respectively. Likewise, in the simulations results from the data collected from the network planning tool in the rural environment, PCA-KDE methods can achieve significant improvement in both partition models. In particular, the 50th percentile of the distance estimation errors when using Cluster-based PCA-KDE is around 39.5 m, by using Grid-based PCA-KDE are around 57.4 m. Although the data set from the simulated propagation model and network planning tool seems too simplified without considering the complex environmental factors and fading factors, it also can be concluded that using

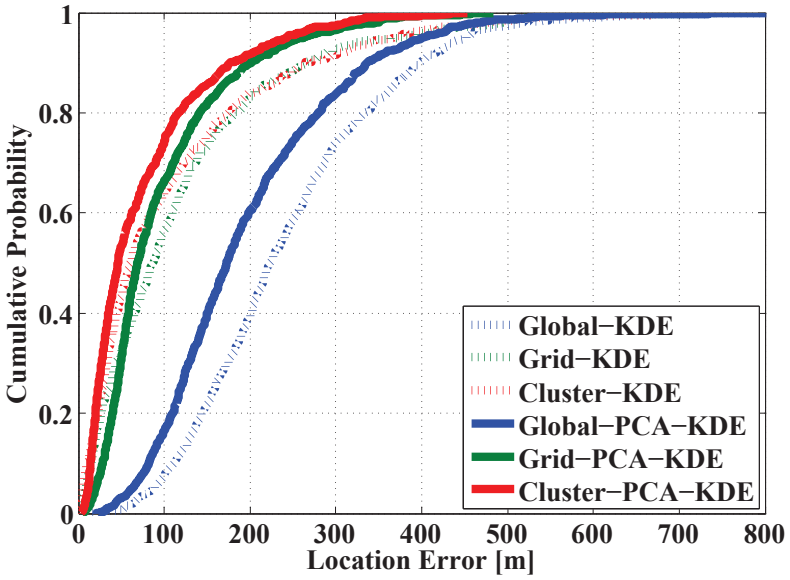
PCA to reduce the correlation between signal strength before using probability methods to build RSS distribution model for positioning are better than directly using probability methods by assuming signal strength independently. In the orthogonal space generated by the construction of the principal components the simple multiplication to compute the joint probability is valid, whereas in the non-transformed space it is not.

6.5.1.2 Location Results with Real Data Sets

Figure 6.4 and Table 6.4 illustrate the comparison results using the real data sets from the two different outside environments. For the data set from Queen Mary campus (see Figure 6.4 (a)), it is obvious seen that KDE and PCA-KDE methods based on cluster model outperform these two methods based on grid model and global model. More specifically, for the cluster-based PCA-KDE method, the mean values of the distance errors are within 85.5 m, and KDE method based on clustering scheme reports 128.1 m in mean value. What's more, the 50th percentile of the distance estimation errors when using Cluster-based PCA-KDE is around 66.4 m. Likewise, for the Music Festival in Victoria Park, the PCA-KDE based on the cluster model also performs slightly better than the result of the other three methods. Unlike outdoor scenario 3, it can be seen that the positioning accuracy performance of the two algorithms with respect to partitioning model, the cluster model is quite similar to the grid model. This is primarily because of the few BSs located nearby. In summary, applying PCA into KDE algorithm can improve the estimation accuracy in the real outdoor environments tested.



(a) Outdoor Scenario 3: Queen Mary campus



(b) Outdoor Scenario 4: Music Festival in London Victoria Park

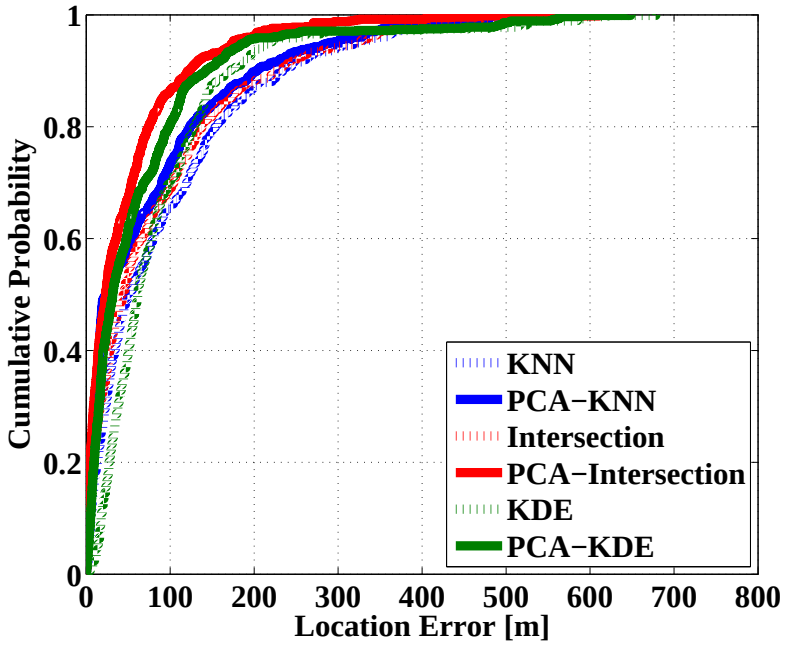
Figure 6.4: Cumulative percentile of error for KDE and PCA-KDE based on different partitioning models in simulated environments: (a) A Simple Simulated Urban Propagation Model; (b) Music Festival in London Victoria Park

Table 6.4: Comparison of Estimation Error between KDE and PCA-KDE Methods based on Global, Grid and Cluster Models in Outdoor Scenario 3 and 4 (in meters)

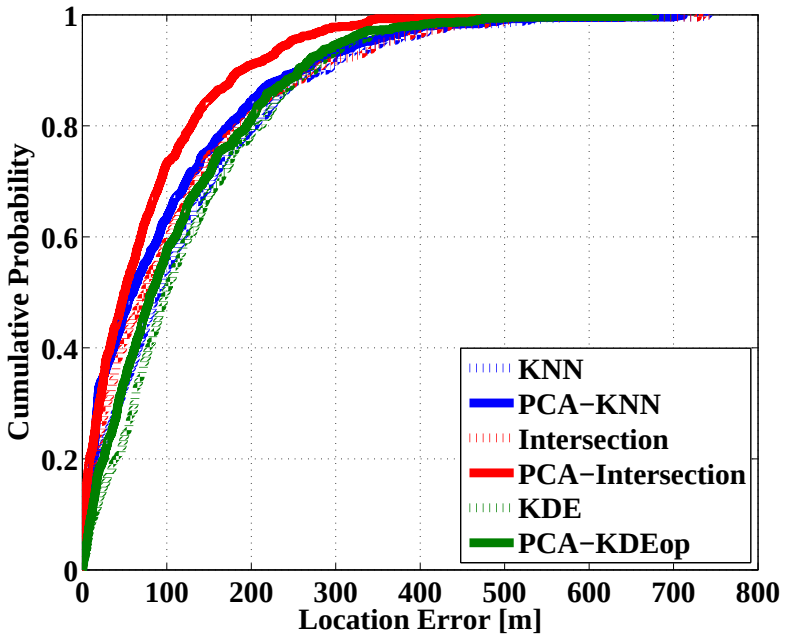
	Outdoor Scenario 3: Queen Mary campus					
	KDE			PCA-KDE		
	Global	Grid	Cluster	Global	Grid	Cluster
Mean Error	210.8	128.1	112.0	173.9	113.7	85.5
50 Percentile	191.8	80.0	90.3	156.5	72.7	66.4
75 Percentile	249.3	160.7	130.8	204.9	130.8	89.6
90 Percentile	322.8	273.2	189.5	271.7	255.1	141.5
	Outdoor Scenario 4: Music Festival in London Victoria Park					
	KDE			PCA-KDE		
	Global	Grid	Cluster	Global	Grid	Cluster
Mean Error	243.1	118.1	108.0	198.0	95.9	76.1
50 Percentile	222.0	86.0	59.9	171.7	69.0	46.3
75 Percentile	308.6	155.8	147.7	258.2	124.0	103.2
90 Percentile	396.0	264.8	263.9	344.2	203.6	180.0

6.5.2 Augmenting PCA to Improve Accuracy

Three different algorithms, viz. Intersection, KDE and KNN, each with and without the use of PCA, are used to evaluate the benefits of PCA in a real environment. For KNN, the constant K is set as 3 in this experiment. Figure 6.5 depicts the CDF of the error distance for the KNN, Intersection and KDE algorithms with and without PCA in both Queen Mary campus data set and the first day data set in London Victoria Music Festival. In comparison, the figure clearly shows that the proposed approaches outperform the traditional approaches. More specifically, for Queen Mary campus (see Figure 6.5 (a)), using the clustering scheme, the percentiles within 50 meters for the original KNN, Intersection, and KDE methods are 49.9%, 54.1%, and 40.9%, whereas for the PCA-KNN, PCA-Intersection and PCA-KDE report 59.6%, 68.0%, and 67.6%



(a) Queen Mary campus



(b) Music Festival in London Victoria Park

Figure 6.5: Cumulative percentile of error for different algorithms with or without PCA in real environments: (a) Queen Mary campus; (b) Music Festival in London Victoria Park

respectively. Hence, it can conclude that using PCA to map the independent axes space performs much better than the original RSS space in a city environment. In addition, as mentioned before, PCA is applied in each cluster. The subsequent application of PCA gives further data reduction (e.g. to 2 or 3 or 4 depending on the cluster) and importantly gives orthogonal axes to support efficient joint probability density function estimation. The transformation matrices are also very different between clusters. Table 6.5 illustrates some of the samples of transformation matrices in different clusters in Queen Mary campus scenario.

Table 6.5: Transformation Matrices in Different Clusters in Queen Mary Campus Scenario

Cluster ID	Transformation Matrices
Cluster 2	$\begin{bmatrix} 0.5021 & -0.5223 & 0.4207 & -0.5459 \\ -0.5619 & -0.4225 & -0.5043 & -0.5014 \\ -0.5257 & 0.4139 & 0.6304 & -0.3937 \\ 0.3946 & 0.6143 & -0.4138 & -0.5437 \end{bmatrix}$
Cluster 5	$\begin{bmatrix} 0.6773 & -0.2240 & 0.7008 \\ -0.2653 & -0.9628 & 0.0513 \\ 0.6862 & 0.1512 & -0.7115 \end{bmatrix}$
Cluster 10	$\begin{bmatrix} -0.7168 & -0.6973 \\ -0.6973 & 0.7168 \end{bmatrix}$

Unlike the city environment in Queen Mary Campus scenario, in the park there are few obstacles, such as high buildings for radio reflections, and the RSS is more closely related to distance. Correlations from different BSs are observed to be lower than that in the city campus environment. Also in a rural area, the BSs are sparse and the RSS values are relatively low. Thus, fewer clusters are needed, and greater estimation errors are reported and the difference between the positioning accuracy of the methods is less.

6.5.3 Reduction in Training Samples Required for Specified Accuracy

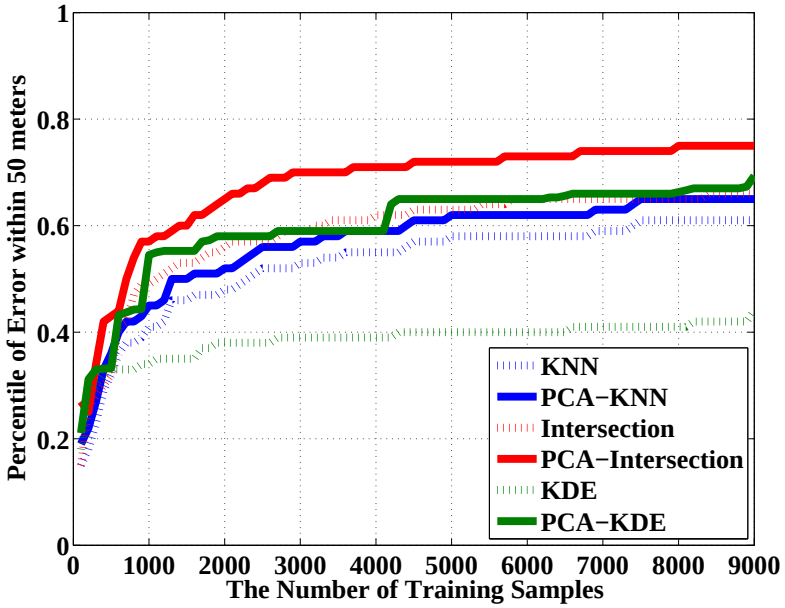


Figure 6.6: Percentile of errors within 50 meters versus the number of training samples in outdoor scenario: Queen Mary campus

The training phase consists of collecting training samples for the radio map. In the proposed method used in the test-beds, the location estimation accuracy does not depend on the size of the area of the interest, but depends on the number of training tuples collected during the training phase: the smaller the number of training data points, the lower the positioning accuracy that is obtained, but the less time the training period takes. A larger number of training data points leads to higher accuracy of location estimation but requires more time for the training procedure. However, the accuracy does not improve in strict proportion to the training sample size and reaches a limit that depends on the algorithm and of course the intrinsic variability associated with the use of RSS. This is why it is useful to have a method that is efficient in its use of the sample data points. Figure 6.6 reports the impact of the number of training samples taken from the Queen Mary campus environment with respect to the estimation accuracy. As seen from Figure 6.6, when only taking 1000 samples, the PCA-Intersection algorithm

already outperforms the other methods, even when these methods use 9000 samples. From Figure 6.6, given a required median accuracy the sample size needs to be chosen to correspond to the Queen Mary campus environment. The results clearly show that the size of the training sample set can be greatly reduced with the integration of PCA into the preferred algorithm for the Queen Mary campus scenario. This can be explained by the ability of PCA to compress more information into fewer dimensions and remove the high correlation between RSSs from different BSs. The extracted PCs provide sufficient information for the model learning, and thus fewer training samples are required in the location system. It can be concluded that applying PCA in the different methods can result in a reduction of costs of site survey and data collection in the specific scenarios considered in this thesis.

6.6 Summary

An improved Kernel Density Estimator method for outdoor fingerprint localisation based on clustering deviation RSS from BSs has been presented in this chapter. Considering the correlation relationship between signal strength, the proposed method applies PCA to obtain a transformed signal strength tuple, so that any two of them are independent. Four different specific scenarios (rural, urban and suburban) have been considered in order to validate the performance of the proposed approach. The comparison results presented show the proposed scheme finds more accurate locations and outperforms the original KDE approach whether based on cluster partitioning or grid partitioning or no partitioning for the specific use cases. Furthermore, in these specific user cases, the results also indicated that applying PCA in Intersection, KDE and traditional KNN algorithms can reduce the computation burden and storage and improve the positioning accuracy. Moreover, fewer training sample were needed in the proposed algorithms when compared to traditional methods in the specific scenarios.

Chapter 7

Outdoor Location Estimation in Changeable Environments

7.1 Introduction

Most previous work assumes the radio map is static. During the training phase, after generating the radio map, location estimation models are built between the RSS and their corresponding location information. Once the models are learnt, they are applied with the radio map for further location estimation without making any adaptation to the new RSS measurements. However, in a changeable environment the observed RSS measurement may significantly deviate from those stored in the radio map due to the changes in humidity, temperature, physical environment, chip set, antenna and the mobile users' hard-to-predict movements. Consequently, location based systems that depend on static radio maps have been criticised because of the often substantial inaccuracies. Therefore, it is a challenging task to design a proper adaptive location estimation approach with respect to dynamic environmental changes.

To take dynamic environmental changes into account, [92] [93] [94] have proposed different approaches utilised inside buildings. By using highly distributed additional

hardware systems, [92] uses a small number of stationary emitters and sniffers to assist location estimation, in order to obtain the new RSS values to update the radio maps in WLAN networks. [93] adapts a static radio map by calibrating new RSS samples at a few known locations and fits a linear function between these values and the corresponding values from the RSS map. [94] applies a model-tree-based method, called LEMT, to adapt radio maps by only using a few reference points in an 801.11b wireless network. LEMT requires building a model tree at each location to capture the global relationship between the RSS values received at various locations and those received at reference points. It needs to install additional sensors to keep on recording RSS values all the time.

In this chapter, a novel algorithm is presented to allow an existing radio map, which is built for a specific weather condition and user population density, to be used under different conditions by using a small set of real-time RSS data points collected for a new weather condition or mobile user density. When applied online the algorithm takes the real-time RSS values and adjusts them so that the existing radio map can be used. This calibration process is not uniform over the area of consideration. The deviations from the existing radio map are clustered and a correction based on the cluster that a run time RSS observation belongs to is applied. It will be seen from the experimental results that this technique can mitigate the variations of signal strength due to different weather conditions and different population densities, and allow calibration without the need to repeatedly rebuild the radio maps for each possible weather condition and user density. The rest of the chapter is organized as follows. Section 7.2 illustrates the proposed algorithm for location estimation using RSS in changeable environment in detail. Section 7.3 analyses the impact of environmental factors on RSS values and presents the experimental evaluation of the proposed algorithm in a real environment. Section 7.4 concludes the chapter and discusses directions for future work.

7.2 Location Estimation in Changeable Environment

To allow adjustment for a changeable environment, the two step localisation workflow (viz. training and online estimation) used above has the second step augmented with a calibration process.

RSS with corresponding location data that are collected at different random locations for a certain environment is called the training data set (TR). So normally there could be several TRs, and one of them will be labelled as the reference (RTR). Typically the RTR could be a large training set, and the other training data sets, called secondary training sets, could be much smaller. In this thesis, only one reference training set and one secondary training set are considered at a time. Firstly the radio map of the RTR is created using clustering and regression techniques as described previously. Secondly, under the desired different weather condition or mobile user density, a small set of the secondary training data (STR) is collected and used to build updating patterns (to be described). Finally, newly measured RSS values are calibrated based on the updating patterns, so they can be regarded as being measured under the reference condition. Hence they can be used for positioning using the proposed methods described in section 7.3.

7.2.1 Training Phase

Reference training data set (RTR):

A set of MSs is collected in the reference environment T_0 in the area of interest. For the j-th RTR element, let $\vec{r}_j(T_0) = (r_{j,1}(T_0), ..., r_{j,q}(T_0))$ represents the signal strength vector received by the MS from q antennas, i.e. BSs and RSs. $\vec{l}_j(T_0)$ represents its corresponding geographic location.

Secondary training data set (STR):

Let n be the total number of data elements in the STR that are measured in environment T_σ. Let $R(T_\sigma) = \{\vec{r}_1(T_\sigma), ..., \vec{r}_i(T_\sigma), ..., \vec{r}_n(T_\sigma)\}$ denote the RSS measurements from nearby transmitters, where $\vec{r}_i(T_\sigma) = (r_{i,1}(T_\sigma), ..., r_{i,q}(T_\sigma))$ is a q-dimension vector of RSS received by STR element i (i.e. a MS) from q antennas. $L(T_\sigma) = \{\vec{l}_1(T_\sigma), ..., \vec{l}_i(T_\sigma), ..., \vec{l}_n(T_\sigma)\}$ consists of the geographic locations. $\vec{l}_i(T_\sigma)$ is the 2-D position coordinates of STR i.

Let the superscript $'$ of r denote the adjusted RSS data. For the i-th element of the STR, its measured RSS values $\vec{r}_i(T_\sigma)$ are adjusted to create $\vec{r}'(T_0)$, so that the estimated signal strength values $\vec{r}'(T_0)$ can be treated as if it was collected in the reference environment T_0.

Step 1: Find the K (e.g. $K = 3$) nearest neighbours of STR i from RTR in location space (not RSS space), and the IDs of these neighbour RTRs are recorded in set U_i. So the physical location of the k-th ($1 \leq k \leq K$) neighbour can be given as $\vec{l}_{U_i(k)}(T_0)$, and $\vec{r}_{U_i(k)}(T_0)$ denotes its corresponding RSS measurements. Therefore, the location distance between STR i and its k-th neighbour can be give as

$$d_k = \left\| \vec{l}_i(T_\sigma) - \vec{l}_{U_i(k)}(T_0) \right\| \tag{7.1}$$

Step 2: Calculate an estimated RSS values for STR i that can be regarded as measured in the reference environment T_0. This can be expressed as

$$\vec{r}'_i(T_0) = \sum_{k=1}^{K} w_k \vec{r}_{U_i(k)}(T_0) \tag{7.2}$$

where w_k is a normalized weight for the k-th neighbour:

$$w_k = \frac{1}{d_k \sum_{i=1}^{K} \frac{1}{d_i}} \tag{7.3}$$

Step 3: Obtain a vector of difference values of STR i between its estimated RSS values $\vec{r}'_i(T_0)$ and measured RSS values $\vec{r}_i(T_\sigma)$.

$$\vec{\Delta}_i = \vec{r}'_i(T_0) - \vec{r}_i(T_\sigma) \tag{7.4}$$

Step 4: Repeat **Step 1-3** for another $(n-1)$ times for all the other data points in STR. Hence every STR has a vector of difference values.

Step 5: Apply the clustering scheme to cluster the n difference values. Let G_i stand for the cluster which $\vec{\Delta}_i$ belongs to. Assume that G_i contains N_i vectors of difference values including $\vec{\Delta}_i$, so the average of all the difference vectors in cluster G_i can be assigned to STR i $(1 \leq i \leq n)$ as:

$$\bar{\Delta}_i = \frac{1}{N_i} \sum_j \vec{\Delta}_j, \left\{ 1 \leq j \leq n \mid \vec{\Delta}_j \in G_i \right\} \tag{7.5}$$

7.2.2 Online Location Estimation Phase

During the online phase for the environment T_σ, given a new MS m with observed RSS tuple $\vec{r}_m(T_\sigma)$ from q BSs, the process of estimating MS m's location $\hat{l}_m$ is as follows.

Step 1: Find MS m's $K^{'}$ (e.g. $K^{'} = 3$) nearest neighbours in the STR (using Eq. (7.6)). Let V be the set that stores the IDs of these neighbouring STR elements, so $\vec{r}_{V(k')}(T_\sigma)$ and $\vec{l}_{V(k')}(T_\sigma)$ can denote the RSS sets and locations of the k'-th neighbour STR respectively. By using the Mahalanobis distance in signal space, the similarity between the MS m's RSS values and its k'-th neighbour STR's RSS values can be obtained:

$$s_{k'} = \sqrt{\left(\vec{r}_m(T_\sigma) - \vec{r}_{V(k')}(T_\sigma)\right)^T \sum^{-1} \left(\vec{r}_m(T_\sigma) - \vec{r}_{V(k')}(T_\sigma)\right)} \tag{7.6}$$

Here $\sum$ is a $q \times q$ covariance matrix in signal space.

Step 2: Based on the similarity in signal space, each of these $K^{'}$ STR neighbours can be assigned a weight, which is defined as:

$$w_{k'}^{'} = \frac{1}{s_{k'} \sum_{j=1}^{K'} \frac{1}{s_j}} \tag{7.7}$$

Step 3: Calibrate the RSS value tuple of MS m to what it would be if it was measured in the reference environment T_0 by

$$\vec{r}'_m(T_0) = \vec{r}'_m(T_\sigma) + \sum_{k'=1}^{K'} w_{k'}^{'} \bar{\Delta}_{V(k')} \tag{7.8}$$

Since the above calibration process focuses on eliminating the impact of environmental factors, such as weather condition and mobile population density, the calibrated RSS value $\vec{r}'_m(T_0)$ can be regarded as measured in the same environment T_0 as reference training data. So the calibrated RSS value can be used for position estimation with the proposed approaches described in chapters 5 and 6.

7.3 Performance Evaluation

Concerning the data collected from a three-day music festival held in London Victoria Park, the weather and population density information during these three days is shown in Table 7.1 according to [95]. Due to the different activities and venues of the music festival, the walking paths on these different days are different, as shown in Figure 7.1.

Table 7.1: Environment Information during the Three Days in London Victoria Park

Day	Temperature	Humidity	Cloud Amount	Precipitation Amount	The Overall Number of People
Day 1	20°C	73%	42%	0.3mm	10,000
Day 2	19°C	64%	54%	0.0mm	30,000
Day 3	16°C	77%	84%	1.3mm	9,000

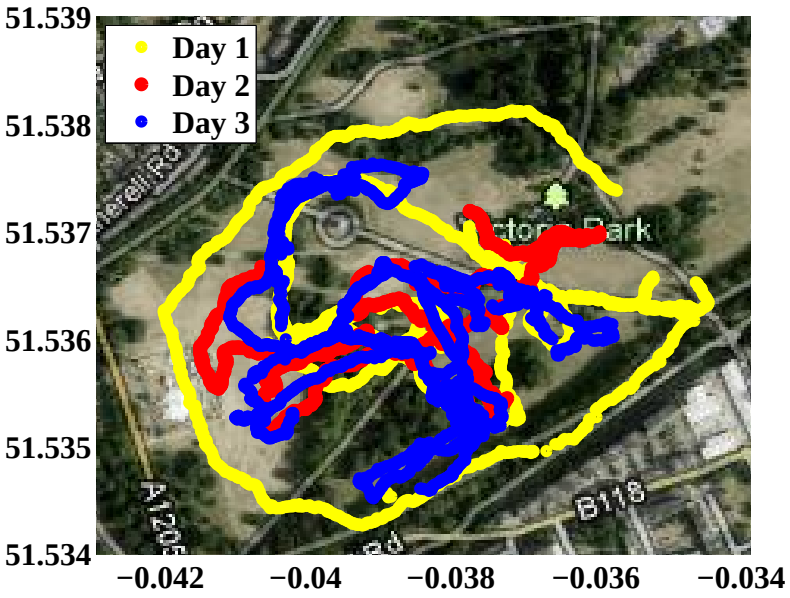


Figure 7.1: Walking paths for each day of music festival

7.3.1 Impact of Changing Environment

The changes in RSS for different conditions are illustrated in the graphs below.

(a) Similar weather, different population density

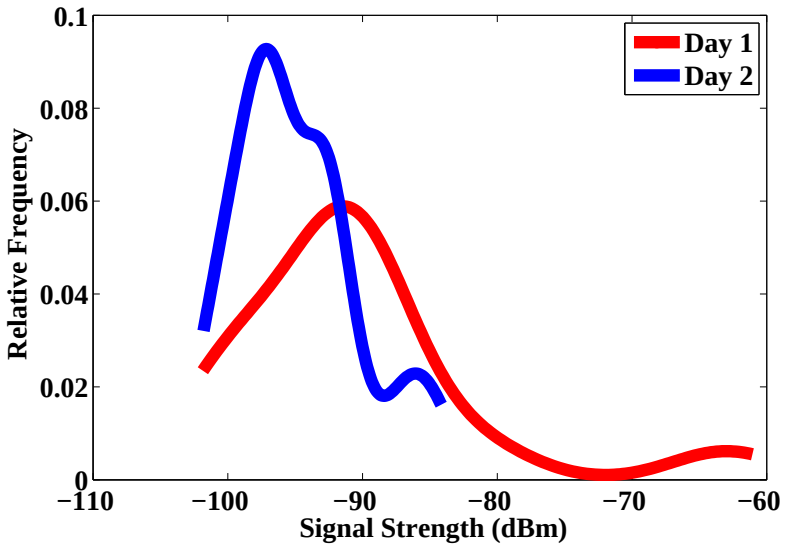


Figure 7.2: The comparisons of RSS distributions for Day 1 (medium attendance) and Day 2 (large attendance) at fixed locations from a typical BS. (Similar weather)

(b) Different weather, similar population density

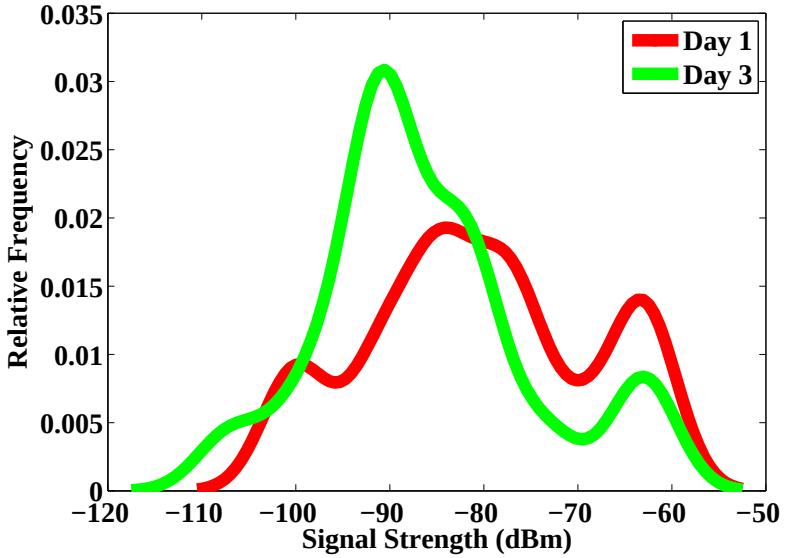


Figure 7.3: The comparisons of RSS distributions over Day 1 (dry and sunny) and Day 3 (wet) at fixed locations from a typical BS. (Similar population density)

It can be seen that the signal strength values received from the same BS at a fixed location may vary significantly. For example, the RSS data points collected during a three-day music festival in London Victoria Park described in chapter 4 demonstrates this effect. From Table 7.1, it shows that Day 1 was sunny and dry and with a moderate number of visitors, while Day 2 had the same weather condition but had a much larger audience. Day 3 was cloudy and wet and with a slight drop in the user numbers compared with Day 1. However, because of the different activities and venue layouts of the music festival on different days, there are only a few locations that are measured with RSS and same GPS signals in all the three days. Therefore, pair wise comparisons of the RSS distributions between Day 1 and Day 2, and Day 1 and Day 3, are made to analyse the impact of environmental factors, e.g. weather condition and population density, on RSS measurements in two cases.

Figure 7.2 and Figure 7.3 illustrate two comparisons of RSS distributions, both of which are processed with the kernel density estimate method. RSS data points in each comparison are measured at the same locations from the same BS. It can be observed that

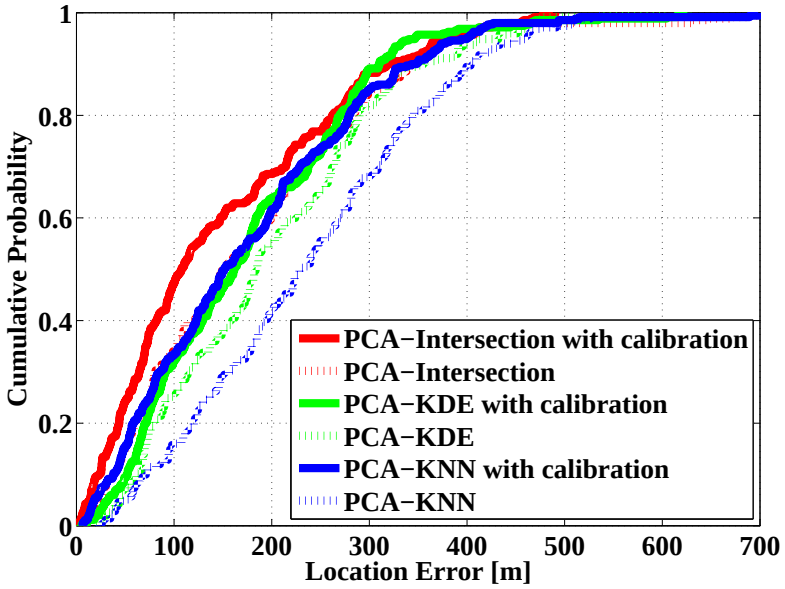
in each figure the signal strength of the peaks vary from each other, probably because of the different environments in each comparison. It can be concluded that the RSS distributions from the same BS vary both with different audience numbers, and weather conditions even at the fixed locations. These variations imply that depending on the original radio map generated in the training phase, the position estimation results might be inaccurate when the physical environment changes.

A network operator may change the transmission power of the BS based on environmental conditions. That is to say, from one day to the next, the transmission power could be different, as well as the environmental conditions. In the proposed localisation scheme as described in Chapter 4, one of the reasons for using the deviations RSS to create clusters is to make the clustering more invariant to the transmission power. The transmission power is a variable that can be adjusted to change the coverage if required. Therefore, the changes of transmission power of the BS do not impact on the estimation accuracy.

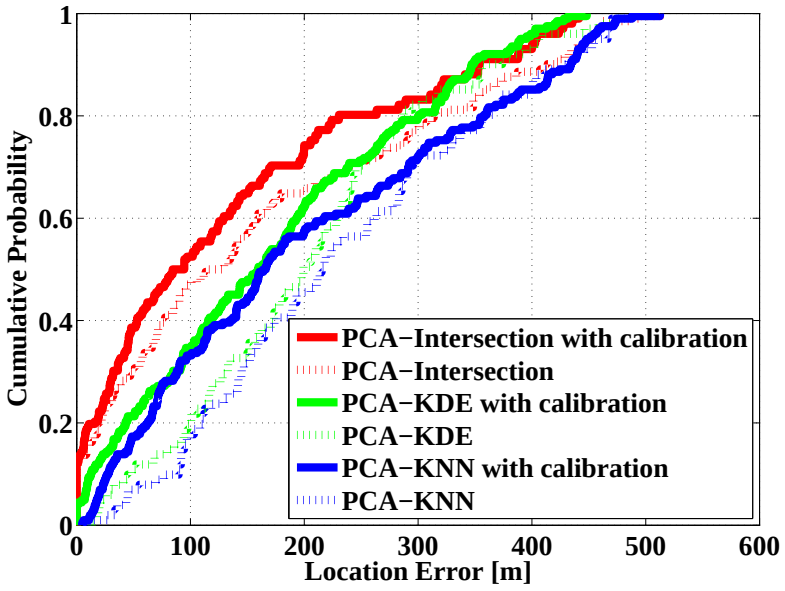
7.3.2 Positioning Performance

There are 2095 RSS samples collected on the first day. 2050 and 3424 RSS measurements with their location coordinates are also collected on the second day and the third day respectively. In each of the two data sets, 600 RSS measurements are randomly chosen as the secondary training data set, while setting the first day as the reference environment, to create the updating pattern. The remaining parts of the measurements are used for location estimation with their RSS values only.

Figure 7.4 depicts the CDF of the error distance for the PCA-Intersection, PCA-KDE and PCA-KNN with and without using the updating scheme on two different days. Comparison of the two figures clearly shows that the data update is effective. More specifically, as seen from Figure 7.4 (a), the percentage of errors less than 150 meters in the PCA-Intersection, PCA-KDE and PCA-KNN methods based on the adapted RSS



(a) Day 2



(b) Day 3

Figure 7.4: Cumulative percentile of error for different algorithms for two different days at Victoria Park Music Festival

data report 60.6%, 46.3% and 49.7% respectively, whereas these three methods without calibrating the RSS are 45.7%, 38% and 29.1% respectively on Day 2. Similarly, it can be observed from Figure 7.4 (b) that the proposed methods with calibration can perform better than those methods without applying any correction for the changes in environment on Day 3, e.g. the mean measurement error of PCA-Intersection with calibration is around 133.2 m, while the PCA-Intersection without any calibration reports 163.9 m.

From the experiments, it can be concluded that the proposed method can make adaptations for changeable environments, giving better localisation accuracy than static fingerprint-based positioning methods in some specific use cases.

7.4 Summary

In this chapter, a novel RSS-based outdoor location estimation approach that can create calibrations so that localisation can be made using the primary radio map has been proposed. This method only needs one full radio map built for a specific environmental condition or user population density and a small set of data points measured in a new environment. The calibrated RSS data points can be regarded as measured in the same reference environment as the training data set. The improvement in location estimation accuracy is tested, and the results show that the proposed algorithms achieve a considerable advantage over previous static fingerprint-based techniques in the three-day music festival held in London Victoria Park.

However, some improvements can be considered: a) in this chapter, 600 samples were randomly selected as secondary training samples from Day 2 and Day 3 respectively, but comparisons with different number of secondary training samples have not been tested yet. Therefore, extensions of this work will focus on how to find the suitable number of secondary training samples in the area of interest; b) this chapter only considered one area with three different scenarios. To validate the applicability of the proposed method,

it is important to collect data for different environments in different areas.

Chapter 8

Network Monitoring with Clustering

8.1 Introduction

Many radio coverage prediction methods have been proposed and typically take one of two approaches [10]: an empirical approach and a site-specific approach. A common problem with empirical models is accuracy, while with site-specific models it is computational efficiency. For empirical models the accuracy of the prediction model is mainly based on the precision of the database used, such as the scope of the database of signal strength measurements and the related topographical information. In the conventional network planning phase, the engineer uses data with location information to identify the radio coverage or link budget. The data is collected by the engineers with special devices on the person or in a vehicle. Although there are some other tests as part of the network state information collection, RSS is the main parameter for the radio coverage estimation and also applied in the approach described here.

It is too difficult to describe an area using only one model because of the complexity of the propagation environment. Previous research has relied on the mobile users' exact

location in order to calculate the prediction of the radio coverage. The propagation model is set up according to propagation theories. Hence, the estimated propagation loss measurement can be obtained. However, it is difficult to ensure accuracy as a high-precision database needs to be created incorporating topographical and build features of the propagation environment. This activity is extremely time-consuming and error-prone. In the approach described in this chapter the exact locations are not necessary. Clustering allows us to partition the RSS space. Identification of the correct cluster yields a probability distribution of RSS in that cluster. The approach developed in this chapter also allows for assessment of changes in the coverage, arising for example by a new building.

This chapter proposes a mechanism for real-time modelling of the radio coverage in cellular networks so that resource allocation algorithms can benefit from traffic demand distribution and coverage information. This gives fundamental support for the self-processes (configuration, optimisation) in SON. In this work, a large outside area is partitioned into small clusters created by analysis of RSS data points collected from historical data, i.e. during a training phase, based on the proposed clustering scheme as introduced in chapter 4 and monitoring the current mobile users' RSS to access the current radio coverage status. With this knowledge, novel optimal configurations for the antennas, e.g. power, tilt, interference control and the frequency allocation, can be selected with better assurance that the prospective network performance on the real terrain will be adequate and, hence, achieve a better QoS. Therefore, in this section the real time monitoring of radio coverage is explored.

The main contribution of this chapter is twofold: 1) the proposed approach models the probability density of RSS in every small area partitioned by clustering, rather than constructing a propagation model to predict the received power for a given location inaccurately. This is better to represent the reliability; 2) the proposed clustering provides a good support for estimation of radio coverage in complex outdoor environment. The potential exploration of this approach could also be used to assess the network state and

affected civilians in emergency or disaster situations.

The remaining part of the chapter is organized as follows: section 8.2 presents the proposed run-time self-training measurement system and the corresponding approaches for predicting the radio coverage and adapting to a dynamic environment are presented in detail in section 8.3 and section 8.4 respectively. The performance evaluations of the proposed algorithm are performed in section 8.5 over data points generated from a network planning tool in a real environment. Finally, section 8.6 discusses the results and outlines open issues for future research.

8.2 The Overview of Run-time Self-training Measurement Mechanism for Coverage Prediction

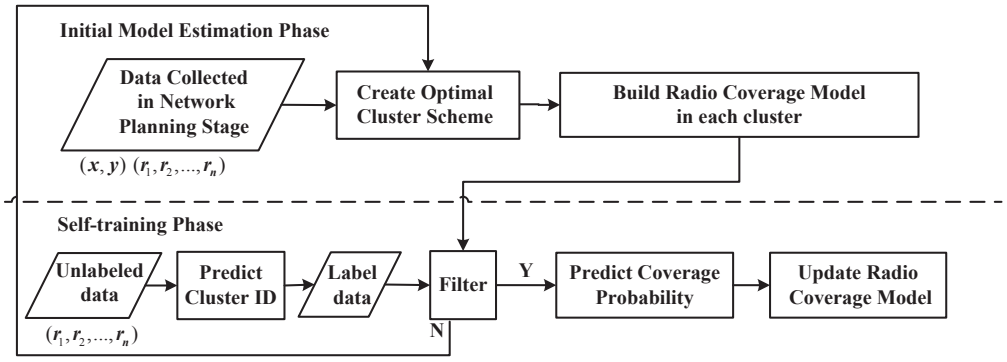


Figure 8.1: Illustration of the run-time self-training measurement mechanism for coverage prediction.

The proposed run-time self-training measurement mechanism for coverage estimation involves two phases: the initial model estimation phase (a.k.a. the training phase) and the self-training phase (a.k.a. the online phase), which is illustrated in Figure 8.1. This mechanism can be integrated with the one introduced in chapter 4 (see Figure 4.1) to be a comprehensive mechanism in an outdoor environment. The black arrows in this figure show direction of flow from one step to another.

8.2.1 The initial model estimation phase

Finding the clustering scheme and radio coverage models are the main aims in the initial model estimation phase. Chapter 4 described the proposed clustering scheme to partitioning the environment into different disjoint regions where each region in RSS space maps to locations in the real environment that have similar RSS. This is achieved by creating clusters in the space of RSS. The experiment results using real data sets in chapter 4 has shown that the proposed clustering scheme partitions the environment into consistent geographic regions that are more homogenously covered by the radio signal, and also the geographical distribution of the created clusters reflected better the RSS distribution and better model the realistic environment according to the RF propagation. For each partition of RSS space, accurate radio coverage models are built to ensure the reliability of coverage prediction in the self-training phase.

8.2.2 The self-training phase

In the self-training phase, the objective is to find novel radio coverage models, which can ensure network performance on the real terrain and achieve a better QoS in a dynamic environment. To account for the complex propagation environment found in a real wireless network, the prediction model is initialized with the training data points first, which can be extracted either from the data from the real environment during an experimental training period or from simulation models. Thus, in the coverage prediction domain, the purpose is to focus on how to use the real-time mobile users' RSS with self-training learning to improve the accuracy of the pdfs of the RSS, in response to the antenna re-configurations, and detect changes in the environment.

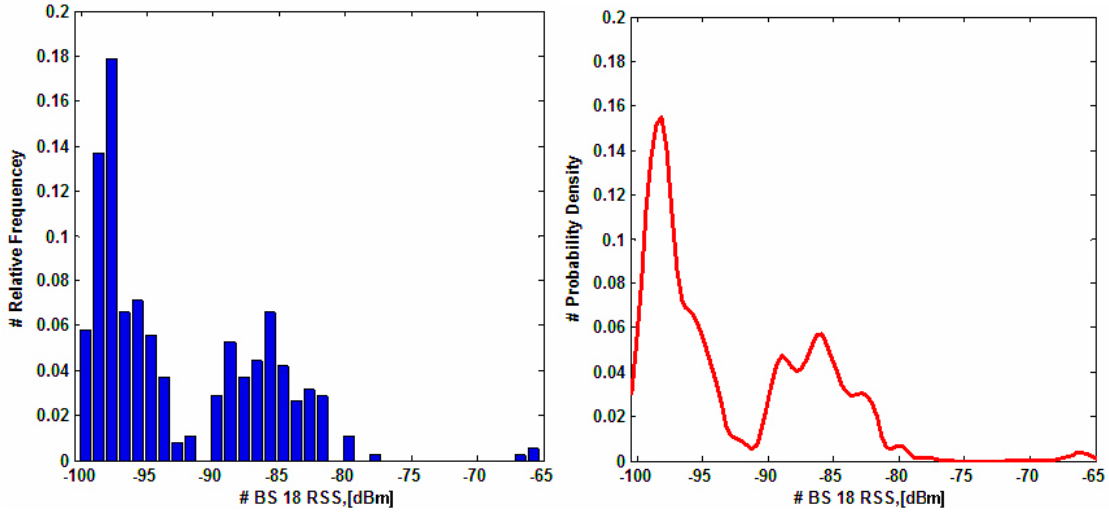
The self-training learning process is based on a *semi-supervised* learning technique. The motivation for having a self-training process stems from the use of unlabeled data to help build a better classifier from the labelled data points. At each time interval (e.g. one hour), when a set of new MSs with observed RSS tuples has been collected, the best

matching cluster is found for each new MS according to their respective received power by using the K-Nearest Neighbour-Venn Probability Machine (KNN-VPM) described in chapter 4. The algorithm only needs to find which cluster the mobile user belongs instead of calculating its exact location. In this way, the algorithm is tolerant to a location calculation error. In order to take physical environmental changes into account, a filtering model is employed to detect if there are environmental changes in the surrounding area. The main idea is to create RSS coverage models and tests for discrepancies in the coverage models created from additional data points that can be collected periodically. If there are statistically significant differences between the historical model and the additional model, the system will re-cluster the RSS in this area and build a new coverage model. If there is no (significant) difference, the self-training learning scheme can reuse the previous model rather than re-creating a new model. This will be presented in section 8.4. Moreover, since the coverage prediction model is relatively simple and there is no accurate location requirement, this means the prediction model can be used in an operational phase for further optimisation.

8.3 Coverage Probability Prediction with Clustering

Assume that in the training phase the following data is collected for a set of n MSs: the MS geographic location and the RSS measurements from neighbouring transmitters. Using the clustering scheme, the terrain is divided into a set of clusters $C = \{C_1, C_2, ..., C_N\}$ where N is the total number of clusters. These created clusters construct a radio map, which not only capture the characteristics of the signal propagation in given environments, but also avoid the modelling of the complex radio propagation and can reduce the computational cost of the coverage prediction. If M denotes the radio map, the i-th element in the radio map can be expressed as

$$M_i \quad = \quad (C_i, \{\vec{r}_j = (r_{j,1}, r_{j,2}, ..., r_{j,q}) | j \in n_i\}, \{l_j | j \in n_i\}) \tag{8.1}$$



(a) Histogram of RSS measurement distribution (b) Kernel density estimates of RSS

Figure 8.2: The comparisons of RSS distributions for one BS by using histogram and Kernel density estimates for one dimension

Where C_i is the i-th cluster ID, n_i is the variable presenting the number of training MSs within cluster C_i. Let $\vec{r}_j = (r_{j,1}, r_{j,2}, ..., r_{j,q})$ represent the set of RSS received by MS j from q antennas, i.e. BSs and RSs, in the area of interest in cluster C_i and l_j the corresponding location of MS j.

Given the radio map, modelling of the coverage in each cluster is undertaken. For complex changes, such as tilt, the reflections will be different and so the modelling is intended to capture the changes at a statistical level, such as the probability that the RSS will exceed a threshold at different locations. In this thesis, the Kernel Density Estimate (KDE) technique is also adopted instead of the histogram technique to build a coverage radio model in each cluster. Because RSS histograms are discontinuous it is very difficult to capture the structure of the set of observed data, and this problem is even worse with more than one dimension, since it requires a very large amount of data to be sampled or else most of bins would be empty. This is illustrated in Figure 8.2(a). Also there is variability, whatever the estimation technique. The data is also sensitive to changeable weather patterns, vehicle movement and people walking. Using the KDE to model RSS distribution can smooth the discrete histogram to a continuous function

and better illustrate the density estimate of the observed data shown in the red line in Figure 8.2 (b).

In the coverage prediction domain taken in this thesis, the estimation of the distribution of RSS in each cluster is constructed for the strongest signal strength values of the MSs in corresponding cluster. If $r_{i,max}$ is the strongest RSS value obtained by MS i from all the neighbouring BSs, the Gaussian KDE of the unknown density in each cluster is:

$$\hat{f}(r) = \frac{1}{n}\sum_{i=1}^{n} K(r; r_{i,max}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}\delta_r} \exp(-\frac{(r - r_{i,max})^2}{2\delta_r^2}) \tag{8.2}$$

Here n is the number of training MSs in one cluster, and δ_r is the smoothing parameter that determines the width of the kernel. The optimal δ_r is obtained by minimizing the AMISE between the estimated and true densities according to (6.6). Since in coverage prediction, only the strongest RSS value of every RSS vector are taken in account, $d = 1$, $\delta_r = \{\frac{4}{3n}\}^{\frac{1}{5}}\sigma$.

To evaluate the stability of the coverage model, a certain proportion of the RSS samples are randomly selected for training in each cluster and the rest of data is used for testing. Initially, a small amount of RSS data is used to construct the coverage distribution model, P_{t_0}, during the offline period, and then at every time interval, t_i, this coverage model is refined by adding a small amount of traffic data for each cluster, P_{t_i}. Let P_{t_m} denote the final RSS distribution model constructed from the whole RSS measurements in each cluster. The *Kullback-Leibler distance (KL-distance)* [96] is applied to quantify how close the RSS distribution models built over different time periods are to the final RSS distribution model construct from the whole RSS measurements in each cluster. This distance is given by:

$$d(P_{t_m}, P_{t_i}) = \int_{-\infty}^{+\infty} f_{t_m}(r) \log \frac{f_{t_m}(r)}{f_{t_i}(r)} dr, i = 0, ...m - 1 \tag{8.3}$$

Here $f_{t_i}(r)$ and $f_{t_m}(r)$ are the densities of P_{t_i} and P_{t_m} respectively. The KL-distance

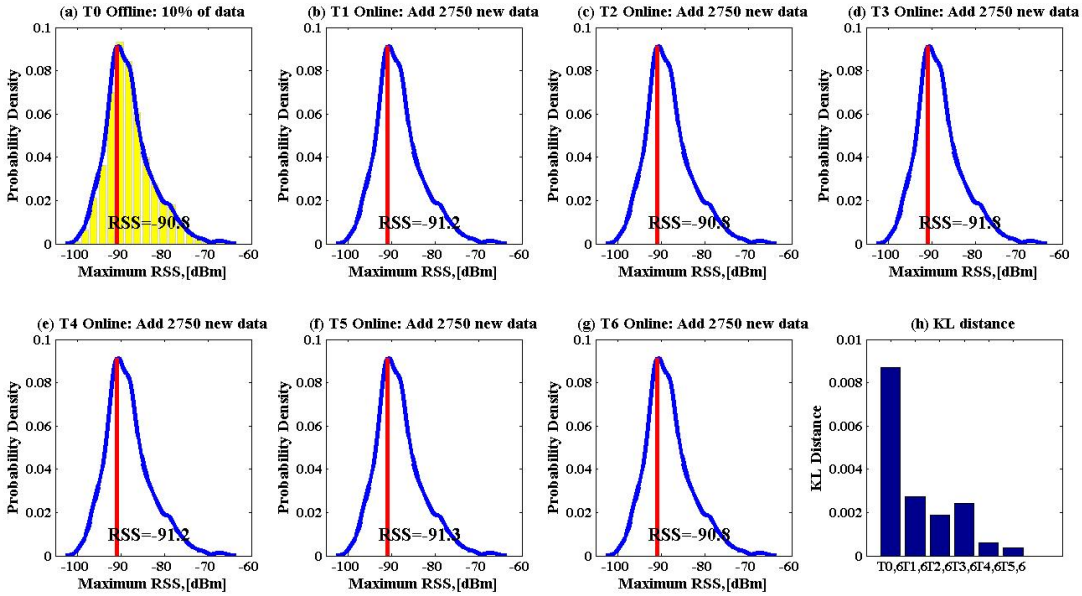


Figure 8.3: The variations of coverage distribution over different time periods in one cluster.

is always non-negative, and a larger KL-distance value from two probability density functions implies a greater difference between them. In Figure 8.3, a typical example is given to illustrate the variations of coverage distribution model over different time periods in one cluster. From Figure 8.3 (a) to Figure 8.3 (g), it can be clearly observed that these RSS distribution models at different time periods are very similar, but not identical. The KL-distance values are in Figure 8.3 (h). Although Figure 8.3 (h) shows the coverage model constructed in the off-line phase has the largest deviation from the coverage model built from the whole RSS measurements, the KL-distance value between them, $T_{0,6}$, is very small and nearly zero. Therefore, it can be concluded that even using a small amount of traffic data can build a fine RSS distribution model in each cluster.

Radio coverage probability is used to reflect the quality of communication in an area. The coverage probability is defined as the probability that the received powers of all possible area elements within the target service area exceeds a specific threshold value. Using the clustering scheme, estimates of radio coverage probability within each cluster based on observed RSS data can be made. Within one typical cluster (i.e. C_1), according

to (8.2) the radio coverage model can be built from the strongest signal strength values of the training MSs in the corresponding cluster. Let $\hat{f}_{C_1}(r)$ denotes the pdf of RSS in cluster C_1. For a chosen threshold $R_{threshold}$, the coverage probability in cluster C_1 can be expressed as

$$P_{C_1}(r > R_{threshold}) = \int_{R_{threshold}}^{+\infty} \hat{f_{C_1}}(r)dr \tag{8.4}$$

8.4 Adapting to a Dynamic Environment

To take dynamic environmental changes into account, a filtering model is proposed to monitor and detect changes in the real environment in future research work. As previously mentioned, in the training phase, the pdf of RSS analysis in radio coverage model provides an initial estimate of the signal characteristics and channel model. So that, during every T time periods in the self-training phase, a new set of RSS observations is collected and these observations are allocated to their best candidate cluster ID using KNN-VPM algorithm. Using the KDE technique, a new coverage model in each cluster is constructed from these new observed RSS measurements without the training RSS samples in corresponding clusters. That is to say, two coverage radio models are built in each cluster: one is based on the RSS measurements during the training phase; the other one is based on the new RSS measurements during the online phase. The *Kolmogorosv-Smirnov test* (K-S test) [97] is applied to compare the RSS distribution models derived from two different traffic samples. If these two coverage models (i.e. distributions) based on training data and new observations in the same cluster differ significantly, the current environment is changing. This could be because of weather conditions (e.g. rain), new buildings erected, etc. That means a re-clustering with the new RSS values is needed as shown in Figure 8.2.

8.5 Simulation Results

To test the efficiency and robustness of the self-training learning method and the accuracy of coverage prediction, the data generated by a network planning tool ASSET 3G for the island of Jersey is used. Six BSs are chosen in the centre of the island covering an area of 8 km $\times$ 6 km. 160 clusters are created using the proposed clustering scheme as previously mentioned in chapter 4 section 4.6.1.2. The distribution of the generated clusters can represent the topographical features significantly, including the contours of highways and roads.

In outdoor environment, the threshold value of radio coverage is usually set to be -90 dBm. In this case, the coverage probability status for each cluster in the central area of Jersey is illustrated in Figure 8.4. For each cluster, if the predicted coverage probability is below 60%, the corresponding cluster area is considered to be an un-served place (see black-coloured area in Figure 8.4). As seen from Figure 8.4, the coverage probabilities for the majority of cluster areas in the central area reach above 90%, and few partitioned regions are un-served. Thus, the clustering scheme gives good support to guarantee high homogenous coverage within a target service area in the central area of Jersey.

In addition, in order to better ensure the reliability of the predicted coverage probability in this test-bed, Figure 8.5 depicts the coverage status for every user in the whole area. Similarly, if the strongest signal power received by one user is smaller than the threshold value, -90 dBm, and this user is considered as outside the served area and marked with black point in the figure. Comparing these two figures, it can be clearly observed that the distribution of estimated coverage probability within clusters is in line with the distribution of the maximum RSS measurement for every user. Taking a cluster as a region to compute the coverage probability, the average of the difference value between estimated coverage probability and measurement coverage probability based on clusters is 1.11%, and for the majority of clusters, the coverage probability bias is nearly 0%, as illustrated in Figure 8.6. These simulation results demonstrate that the proposed

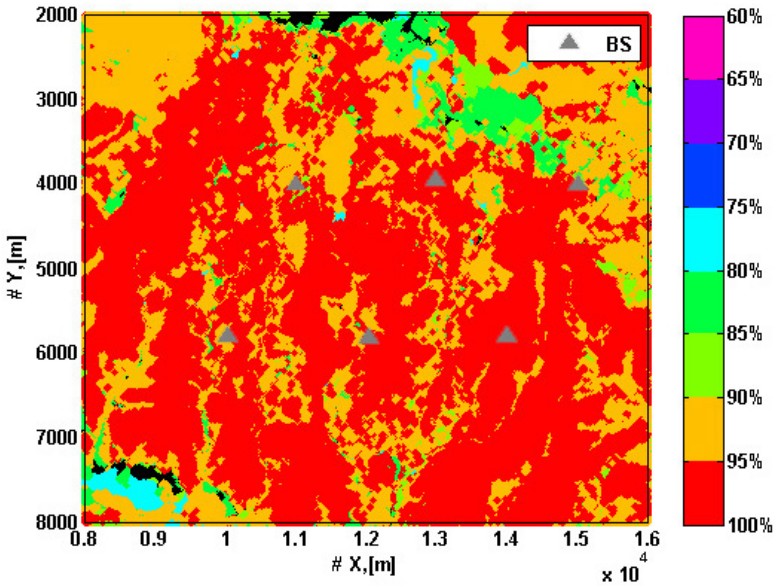


Figure 8.4: Distribution of estimated coverage probability with clustering in the central area of Jersey

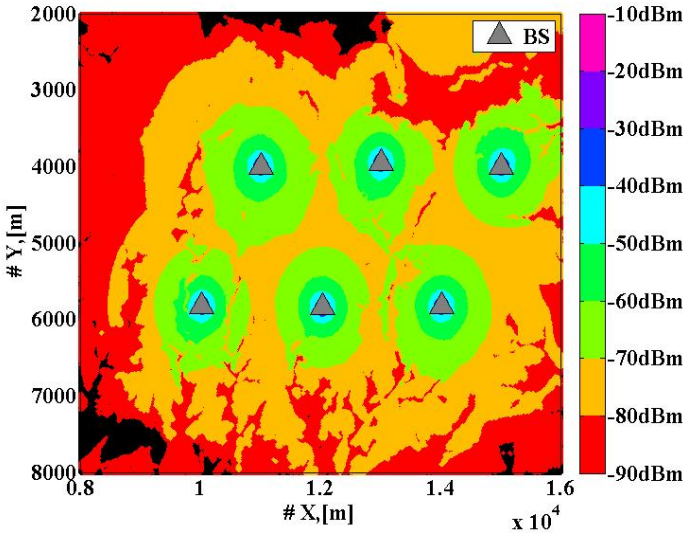


Figure 8.5: Distribution of maximum RSS measurements in the central area of Jersey

radio coverage prediction model based on the proposed clustering scheme in the central area of Jersey is satisfactory and reasonable.

Radio coverage prediction models are dependent on the survey data. Hence, the quality and quantity of the sample data are important for a prediction model. To evaluate

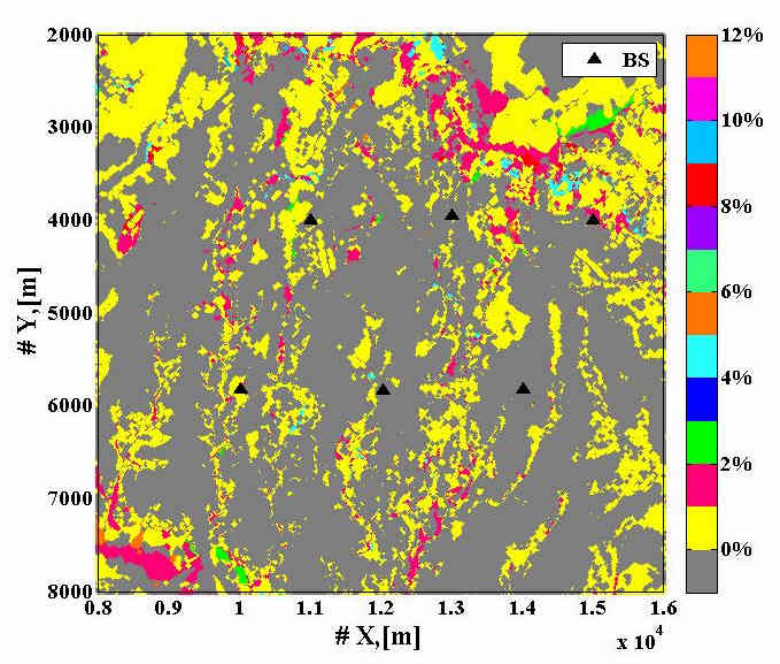


Figure 8.6: The difference between the estimated coverage probability and measurement coverage probability based on clustering in the central area of Jersey

the performance of traffic demand distribution monitoring, the effect of the number of training samples on the performance of the cluster identification method (KNN-VPM) is taken into consideration. In [98], it is reported that there are roughly 88,000 residents in Jersey island with a total area of 116 km^2. Assume the mobile users are uniformly distributed over the whole island, 36,000 mobile users are randomly selected in the central area (48 km^2) and are treated as testing data to do the cluster identification experiment, which represents the accuracy of user traffic distribution estimation. Figure 8.7 displays the estimation accuracy with respect to different number of training data in the central area of Jersey. It can be seen from Figure 8.7 that, initially, the accuracy increases as the number of training data points increases. This is because the more training data points are used, the more information is provided. However, when the number of training data reaches about 50,000, the accuracy of cluster identification is approximately the same. This reflects that there is enough information to distinguish different clusters, and the added training samples cannot contribute to a significant increase in accuracy but to an increase in computational complexity.

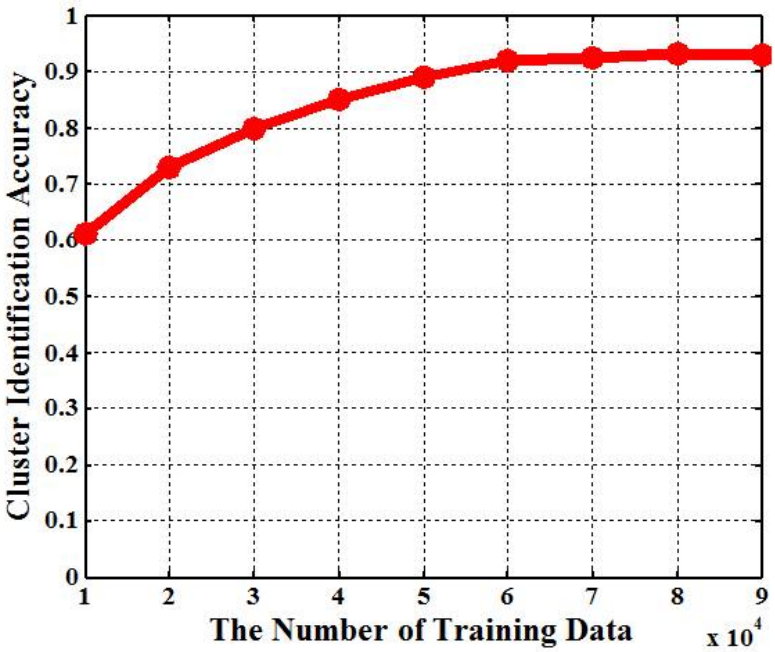


Figure 8.7: The cluster identification accuracy with respect to the number of training data points in the central area of Jersey

8.6 Summary

This chapter has described a run-time self-training measurement mechanism to model the real wireless environment and to determine the coverage probabilities by monitoring of mobile users' RSS in novel configurations. A nonparametric probability approach for modelling radio coverage prediction is proposed to represent and predict the service reliability in a given area based on the pdf of RSS. The self-training semi-supervised learning could remove the need to know the precise location of most of the recorded fingerprints during the training process. By using self-training learning, the model can evolve according to real-time mobile users' RSS values. The traffic distribution for each cluster area can also be roughly estimated based on the cluster identification method. So, the radio coverage and traffic demand could be both monitoring based on the proposed tool. Such models are important for intelligent radio resource management as it allows more accurate hypothetical reasoning and hence the discovery of optimal solutions.

Chapter 9

Location Estimation in an Indoor Environment

9.1 Introduction

As mentioned previously, many localisation systems utilize the signal strength values received from the BSs or RSs or APs to estimate the location of a mobile user, based on deterministic or probabilistic techniques. RSS has been widely investigated principally in the context of indoor location estimation. This is because the data required to create the RSS database is readily collected from indoors. Though not as accurate as time-based methods, RSS fingerprint-based localisation has the potential to overcome the limitations of traditional triangulation approaches, because it performs relatively well for NLOS circumstances where the alternative of modelling the nonlinear and noisy patterns of realistic radio signals is a challenging task. Furthermore, RSS-based methods do not require the cooperation of network operators. The focus in this chapter is on large scale indoor localisation using commercial off-the-shelf (COTS) smartphones. Despite the well-known accuracy limitations of RSS for localisation, this chapter explores methods to provide an accuracy that allows it to be useful for a range of applications. This chapter

also shows that WiFi RSS and GSM RSS data sets can be integrated to enhance indoor estimation accuracy. This has potential to support localisation in certain emergency contexts.

Fingerprinting techniques are especially appropriate for the range of frequencies in which GSM and WiFi networks operate. This is because [99] [100] the signal strength at those frequencies presents an important spatial variability. Regarding GSM technology, several research works use this technology for localisation, especially in outside environment. For example, chapter 4 to chapter 7 utilized GSM-based fingerprinting for outdoor localisation. The RSS fingerprints from the 4-strongest GSM BSs have been collected, achieving 50th percentile accuracy of 29.4 meters in a city environment. While inside buildings, [100] has proposed an accurate GSM-based indoor localisation system by making use of the wide signal-strength fingerprints (includes the 6 strongest GSM cells and readings of up to 32 additional GSM channels, most of which are strong enough to be detected, but too weak to be used for efficient communication), but with the need of dedicated and complex hardware. Many research works [31] [53] [72] have investigated WiFi RSS fingerprinting in a relatively small size of indoor environment for positioning. [53] represents the first fingerprinting system for indoor localisation using portable devices. It localizes a laptop in the hallways of a small office building with accuracies of 2 to 3 meters, using RSS fingerprints from four 802.11 APs. Other work uses additional mechanisms to improve the accuracy of this technique, such as RFID [47] and Zigbee [101]. Methods that use auxiliary active RFID tags have been proposed for high indoor accuracy, but this is not ideal for general use in larger areas.

The aim of this chapter is to provide a novel hybrid RSS-based localisation method for indoor environments and test it in a large indoor multi-floor environment. Unlike other previous research, this work is content to locate to a specific room or store or store segment for a mobile user rather than looking for very high location accuracy indoors. In this work, a hierarchical partitioning scheme is used to divide the MSs in the training set into a tree of clusters according to the sequence of the transmitter labels sorted by their

RSS values in a descending order. For example, the first level branches correspond to partitions where each particular transmitter is the strongest; all MSs with the same two transmitters in the same order of RSS form a second level branch from the root of the tree. At run time, for a new MS with a given RSS tuple, the labels of the transmitters that cover this MS sorted by RSS are used and this can be determined which cluster the MS belongs to, by finding the longest label match in the tree. Then PCA is used to transform the RSS into the transformed basis for that cluster. After transforming, the WKNN algorithm is applied to predict the room number for this MS in that cluster.

Techniques based on radio maps become increasingly inaccurate over time because APs fail, or are turned off, or change. These issues are amplified in emergency contexts as then many APs could fail simultaneously. Solutions in the literature to date require reconstruction of the map. (This is done periodically by Google for outdoors. Even the highest resolution of Google, without using GPS, can be rather inaccurate and can deteriorate with time.) This chapter addresses this issue and extends it to emergency situations in a large indoor environment. The main idea is to figure out the estimated failed AP IDs by the use of GSM data. Naively the GSM data set can be used as a backup if some of APs cannot work, but better the GSM data can be used to retrieve some of the lost WiFi accuracy in areas where coverage is reduced.

The novel contributions of this chapter are: a) this work focuses on a large scale multi-floor shopping mall; b) it takes the different importance for WiFi and GSM signals into account, using clusters based on WiFi RSS and GSM RSS respectively. Hence the estimated room number can be obtained using different weightings for the cluster sets from these two kinds of RSS; c) the experimental results show that integrating WiFi RSS with GSM RSS data points can marginally enhance positioning accuracy; d) in order to improve the estimation accuracy in the emergency situations, GSM data can help to estimate the possible failed APs to update the radio map, with a more marked increase in accuracy.

The rest of the chapter is organized as follows. Section 9.2 illustrates the proposed

algorithm for the room estimation using both GSM RSS and WiFi RSS in indoor multi-floor buildings in detail. Section 9.3 describes how to locate a user in an emergency situation. The experiment environment and the experimental evaluation of the proposed algorithms are presented in section 9.4 and section 9.5. Section 9.6 concludes this chapter.

9.2 Localisation in a Static Indoor Environment

9.2.1 Training Phase

The aim is to build a radio map during the training phase. Let q_1 and q_2 be the number of WiFi APs and GSM BSs respectively. Let $Z = \{z_1, ..., z_i, ..., z_n\}$ be the set of the n training MSs, where $z_i = \left[\vec{r}_i^{wifi}, \vec{r}_i^{gsm}, l_i\right]$, $\vec{r}_i^{wifi}$ and $\vec{r}_i^{gsm}$ are the WiFi RSS tuples and the GSM RSS tuples for the i-th training MS respectively and l_i denotes its corresponding store (or store segment) number.

Algorithm 9.1 Hierarchical Partitioning Scheme in Indoor Training Phase

Required:

$Z = \{z_1, ..., z_i, ..., z_n\}$: training data set

w, g: the number of transmitters to choose from APs and BSs respectively (the matching length)

Steps:

1: **for** $i = 1$ **to** n **do**
2: Sort the strongest w WiFi RSS in descending order and get corresponding ID sequence of APs: $[i_1^W, i_2^W, ..., i_w^W]$ as the ranking ID pattern for i of length w, $\mathcal{P}_i^{W,w}$.
3: Sort the strongest g GSM RSS in descending order and get corresponding ID sequence of BSs: $[i_1^G, i_2^G, ..., i_g^G]$ as the ranking ID pattern for i of length g, $\mathcal{P}_i^{G,g}$.
4: **for** length $d = 1$ **to** w **do**
5: Let $C_i^{W,d}$ be the cluster that z_i with ID pattern $\mathcal{P}_i^{W,d}$ belongs to.
6: $\forall j < i$, **if** $\mathcal{P}_i^{W,d} = \mathcal{P}_j^{W,d}$, **then** $C_i^{W,d} \equiv C_j^{W,d}$
7: **end for**
8: **for** length $d^{'} = 1$ **to** g **do**
9: The ID pattern for z_i in length $d^{'}$ is $\mathcal{P}_i^{G,d^{'}}$.
10: $\forall j < i$, **if** $\mathcal{P}_i^{G,d^{'}} = \mathcal{P}_j^{G,d^{'}}$, **then** $C_i^{G,d^{'}} \equiv C_j^{G,d^{'}}$
11: **end for**
12: **end for**

The training phase analyses the RSS of WiFi and GSM separately with the same procedures which are illustrated in Algorithm (9.1). For WiFi measurements of every

training data z_i (step 1), the strongest w WiFi RSS measurements in descending order are selected, which can be expressed as

$$\left\{ r^{wifi}_{i,i^W_1} \geq r^{wifi}_{i,i^W_2} \geq ... \geq r^{wifi}_{i,i^W_w} \mid i^W_1, i^W_2, ..., i^W_w \in [1, q_1] \right\} \tag{9.1}$$

Here $i^W_1, i^W_2, ..., i^W_w$ are the ID series of the chosen w WiFi transmitters respectively. This series $[i^W_1, i^W_2, ..., i^W_w]$ can be regarded as the order w ID ranking pattern $\mathcal{P}^{W,w}_i$ of z_i for WiFi networks (step 2). Similarly, the corresponding IDs of BSs can be obtained as $\{i^G_1, i^G_2, ..., i^G_g\}$ by choosing the strongest g RSS of GSM transmitters in descending order for z_i (step 3). For WiFi data, by choosing lengths $d \in [1, w]$ for the length of the ID sequence (step 4), the w ID patterns $\mathcal{P}^{W,d}_i$ for z_i can be extracted (step 5). For a specific length, training data points having the same ID pattern can be regarded as belonging to the same cluster (step 6). Similar steps are used to create clusters in the training data from the strongest g GSM RSS (step 8 to step 11).

9.2.2 Location Estimation Phase

Given a new MS m with observed RSS measurement $\vec{r}_m = (\vec{r}^{\,wifi}_m, \vec{r}^{\,gsm}_m)$ from q_1 APs and q_2 BSs, the aim is to estimate which room this MS belong to, which are illustrated in Algorithm (9.2).

First the RSS of WiFi and GSM are sorted in descending order separately, thus obtaining two ID sequences with the length of w and g, respectively, $\mathcal{P}^{W,w}_m$ and $\mathcal{P}^{G,g}_m$ of MS m (step 1 and 2). The following estimation of the room ID based on WiFi RSS and GSM RSS is similar. For WiFi measurements, the prediction of room ID is implemented recursively within the maximum length w (step 3). Starting from length w, if there is any training data that has the same ranking ID pattern of WiFi RSS of length d within $\mathcal{P}^{W,d}_m$, MS m can be regarded as belonging to the cluster of this training data (step 5 and 6). Otherwise, the search continues using a shorter length for a matching ID sequence

Algorithm 9.2 PCA-WKNN Method in Indoor Online Phase

Requried:

$Z = \{z_1, ..., z_i, ..., z_n\}$: training data set

w, g: the maximum number of APs and BSs respectively in an ID sequence (the matching length)

$\mathcal{P}_i^{W,d}$: all the ranking patterns of WiFi training data in every dimension. $(i \in [1,n], d \in [1,w])$

$\mathcal{P}_i^{G,d'}$: all the ranking patterns of GSM training data in every dimension. $(i \in [1,n], d' \in [1,g])$

$\vec{r}_m^{wifi}, \vec{r}_m^{gsm}$: the MS m's RSS measurements from WiFi and GSM networks.

Steps:

1: Sort the strongest w WiFi RSS of m in descending order and get the ID sequence of APs: $\mathcal{P}_m^{W,w} = [m_1^W, m_2^W, ..., m_w^W]$
2: Sort the strongest g GSM RSS of m in descending order and get the ID sequence of BSs: $\mathcal{P}_m^{G,g} = [m_1^G, m_2^G, ..., m_g^G]$
3: **for** length $d = w$ **to** 1 **do**
4: Extract $\mathcal{P}_m^{W,d}$ from $\mathcal{P}_m^{W,w}$
5: **if** $\exists i$ such that $\mathcal{P}_m^{W,d} \equiv \mathcal{P}_i^{W,d}$ **then**
6: $m \in C_i^{W,d}$
7: Apply PCA-WKNN to estimate a room ID $\hat{l}_m^W$ for m using the training data Z in $C_i^{W,d}$.
8: **break**
9: **end if**
10: **end for**
11: **for** length $d' = g$ **to** 1 **do**
12: Extract $\mathcal{P}_m^{G,d'}$ from $\mathcal{P}_m^{G,g}$
13: **if** $\exists i$ such that $\mathcal{P}_m^{G,d'} \equiv \mathcal{P}_i^{G,d'}$ **then**
14: $m \in C_i^{G,d'}$
15: Apply PCA-WKNN to estimate a room ID $\hat{l}_m^G$ for m using the training data Z in $C_i^{G,d'}$.
16: **break**
17: **end if**
18: **end for**
19: **if** $\hat{l}_m^W = \hat{l}_m^G$ **then**
20: This is the chosen room ID for m
21: **else**
22: Apply PCA-WKNN to estimate the room ID for m using the hybrid RSS training data in $C_i^{W,d} \bigcap C_i^{G,d'}$
23: **end if**

(step 3), and MS m's ranking ID pattern $\mathcal{P}_m^{W,d}$ can be obtained by extracting the first d elements from $\mathcal{P}_m^{W,w}$ (step 4). Once the cluster of MS m's WiFi RSS is found, the room ID $\hat{l}_m^W$ of m can be estimated by applying PCA-WKNN method to the training data in this cluster (step 7), and the estimation process using WiFi RSS terminates (step 8). Since successive signal strength samples from the same transmitter are highly correlated, it should take this high correlation into account to enhance the accuracy. Chapter 4 section 4.5 illustrates how to transform RSS into an uncorrelated basis using

PCA. Likewise, by using the GSM RSS of m, an estimated room ID ($\hat{l}_m^G$) can be also obtained through a similar procedure (step 11 to step 18). By this point, the estimated room IDs for m have been obtained. If these two estimates are exactly the same, the room ID of m can be finally determined (step 19 to step 20). Otherwise, the training data shared by the two clusters selected in step 5 and step 13 is extracted and analyzed using PCA-WKNN to provide a final estimate of the room ID of m (step 21 to step 22).

9.3 Localisation in an Emergency Situation

Secondary training data set (**STR**): Let n be the total number of data elements in the STR that are measured in an emergency situation (T_τ) in a part of the target environment. For example, for one experiment environment (the ground floor in London Westfield Stratford shopping mall), the area marked with red line in Figure 9.1 are chosen where the q closest APs are shut down (The detail about the test-bed can be found in section 9.5.5). Let $z_j(T_\tau) = [\bar{r}_j^{wifi}(T_\tau), \bar{r}_j^{gsm}(T_\tau), l_j]$ denote the j-th STR element, where $\bar{r}_j^{wifi}(T_\tau)$ and $\bar{r}_j^{gsm}(T_\tau)$ are represented the q_1-dimension vectors of WiFi RSS measurements from the q_1 APs and q_2-dimension vectors of GSM RSS measurements from the q_2 BSs respectively. l_j is the room number of STR j.

An important issue with localisation mechanisms is their reliability. In some cases, some access points may be disabled because of local power failures, management, upgrades etc. In such cases, the user may be given location information that is incorrect because, e.g. in the algorithm described the matching process has missing APs. When there are isolated failures then correction is easier. To demonstrate the approach contiguous failures of different sizes are chosen. Hence this section will mainly focus on two issues: (1) How to estimate the number of failed APs with their corresponding IDs in an area where there are contiguous failures in Westfield mall. (2) How to predict the correct position in this situation.

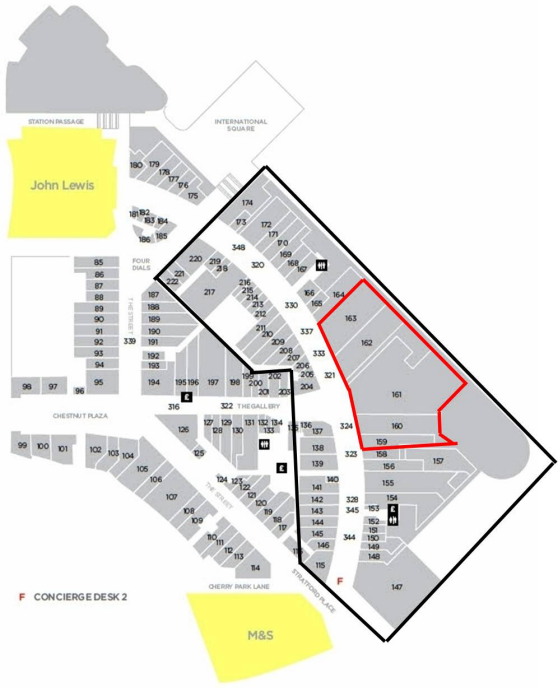


Figure 9.1: A sample of one emergency area on the ground floor in the London Stratford Westfield Shopping mall

9.3.1 Training Phase

In the training phase, suppose that there is a set of MSs collected in the reference environment T_0 in the area of interest where all of APs and BSs in WiFi and GSM networks work well. The MS's room location and its corresponding WiFi RSS and GSM RSS measurements from nearby APs and BSs are recorded. Let q_1 and q_2 be the numbers of WiFi APs and GSM BSs respectively. The collection of this data is taken as **training data set** (**TR**). For the i-th training data, let $z_i(T_0) = [\vec{r}_i^{wifi}(T_0), \vec{r}_i^{gsm}(T_0), l_i]$, where $\vec{r}_i^{wifi}(T_0)$ and $\vec{r}_i^{gsm}(T_0)$ are the WiFi RSS tuples and the GSM RSS tuples for the i-th training MS respectively and l_i denotes its corresponding room number. If one MS does not receive measurable signal strength from one typical AP or BS, its default value is set to -120 dBm, as it is the minimum signal strength. For the TR, the ranking patterns for GSM RSS and WiFi RSS have already been generated according to Algorithm (9.1).

For the j-th element of the STR in the room number l_j, its measured WiFi RSS

values $\vec{r}_j^{wifi}(T_\tau)$ is totally different from the WiFi RSS values of TR data collected in the same room in the reference environment, because of the failed APs. However, the GSM RSS measurements between any two of them in the same room are roughly the same. Therefore, the intention is to estimate the number of disabled APs and their corresponding IDs according to the GSM RSS measurements, in order to update the radio map to improve the estimation accuracy.

Firstly in Algorithm (9.3), the ranking patterns of GSM RSS in TR in different matching lengths are created (Step 1 to Step 7). Then, the estimated number of the failed APs $\hat{q}$ can be obtained (Step 8 to Step 10). Furthermore, for each STR j, it can be assigned to a cluster in TR which has the same GSM ranking pattern with STR j in dimension d (Step 13 to Step 19), so that both the WiFi RSS of the TR data in this cluster and the unique detectable APs' IDs can be obtained. Put these APs' IDs in list B in order to find the most common AP IDs (list U) in this cluster (Step 20 to Step 22). So the possible failed AP IDs can be estimated by comparing the detectable AP IDs list P_j of STR j and the list U, which are added into list T (Step 21). Moreover, traversing all the STR data points, estimated failed AP IDs are regarded as the first $\hat{q}$ AP IDs by counting the number of each possible failed AP ID in list T and sorting them in a descending order (Step 24). Lastly, the WiFi RSS measurements in TR are updated and re-clustered again (Step 25 to Step 32).

9.3.2 Location Estimation Phase

Given a new MS m with observed RSS measurement $\vec{r}_m(T_\tau) = (\vec{r}_m^{wifi}(T_\tau), \vec{r}_m^{gsm}(T_\tau))$ from q_1 APs and q_2 BSs, the estimated room ID this MS belong to can be obtained by using the Algorithm (9.2) based on the updated radio map.

Algorithm 9.3 Updating the WiFi radio map in an emergency situation

Requried:

$Z(T_0) = \{z_1(T_0), ..., z_i(T_0), ..., z_N(T_0)\}$: training data set (TR)

g: the number of BSs to choose from q_2 BSs (the matching length)

$Z(T_\tau) = \{z_1(T_\tau), ..., z_i(T_\tau), ..., z_N(T_\tau)\}$: secondary training data set (STR)

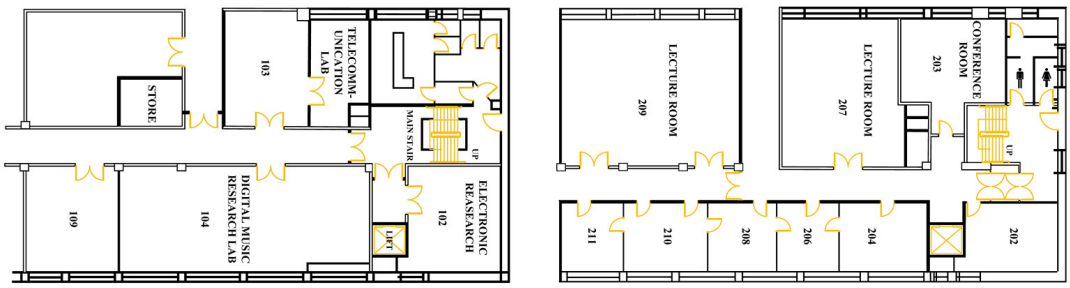
Steps:

for $i = 1$ **to** N **do**
  Sort the strongest g GSM RSS in descending order and get corresponding ID sequence of BSs: $[i_1^G, i_2^G, ..., i_g^G]$ as the GSM ranking ID pattern for TR i of length g, $\mathcal{P}_i^{G,g}$.
  for length $d = 1$ **to** g **do**
    Let $C_i^{G,d}$ be the cluster that TR i with ID pattern $\mathcal{P}_i^{G,d}$ belongs to.
    $\forall k < i$, **if** $\mathcal{P}_i^{G,d} \equiv \mathcal{P}_k^{G,d}$, **then** $C_i^{G,d} \equiv C_k^{G,d}$.
  end for
end for
Calculate the average number of APs detected by every TR i in WiFi networks, LEN.
Calculate the average number of APs detected by every STR j in WiFi networks, L.
The estimated number of failed APs is $\hat{q} = LEN - L$.
for $j = 1$ **to** n **do**
  Obtain the list (P_j) including the detectable AP IDs of STR j according to $\vec{r}_j^{wifi}(T_\tau)$.
  Sort the strongest g GSM RSS of STR j in descending order and get corresponding ID series of BSs: $\mathcal{P}_j^{G,g)} = [j_1^G, j_2^G, ..., j_g^G]$.
  for length $d = g$ **to** 1 **do**
    Extract $\mathcal{P}_j^{G,d'}$ from $\mathcal{P}_j^{G,g}$
    if $\exists i$ in TR such that $\mathcal{P}_j^{G,d} \equiv \mathcal{P}_i^{G,d}$ **then**
      STR $j \in C_i^{G,d}$
    end if
  end for
  Assume there are m TR data in $C_i^{G,d}$, obtain their WiFi RSS measurements and select the unique detectable APs' IDs and put them in a list B.
  if an AP ID in list B can be detected by theses m or $(m-1)$ TR data in $C_i^{G,d}$
  then this AP ID can be taken as one of the most common AP IDs in this cluster and put it in a list U.
  Compare the list (P_j) and the list (U) to find out the possible failed AP IDs, and add these possible failed AP IDs into a list (T).
  Count the number of each possible failed AP IDs in list (T) and sort them in a descending order. The first $\hat{q}$ AP IDs are the estimated failed AP IDs.
  Update the WiFi RSS measurements in TR data and set all the WiFi RSS values from these $\hat{q}$ AP IDs to the value of -120 dBm.
  for $i = 1$ **to** N **do**
    Sort the strongest w updated WiFi RSS in descending order and get corresponding ID sequence of APs: $[i_1^W, i_2^W, ..., i_w^W]$ as the WiFi ranking ID pattern for TR i of length w, $\mathcal{P}_i^{W,w}$.
    for length $d' = 1$ **to** g **do**
      Let $C_i^{W,d'}$ be the cluster that TR i with ID pattern $\mathcal{P}_i^{W,d'}$ belongs to.
      $\forall k < i$, **if** $\mathcal{P}_i^{W,d'} \equiv \mathcal{P}_k^{W,d'}$, **then** $C_i^{W,d'} \equiv C_k^{W,d'}$
    end for
  end for
end for

9.4 Experimental Environment

Two experiments were conducted to evaluate the algorithms. One is two-floors of the EE building in the Queen Mary campus as shown in Figure 9.2; and the other one is the three floors of the London Stratford Westfield Shopping mall as shown in Figure 9.3. For each test-bed, both GSM RSS data and WiFi RSS data have been collected at the same time in each room in the target area by a mobile app on an Android smartphone and their corresponding location information are labelled with the room number or room segment number if the room is large. The downloadable data can be found at [88]. In the Westfield case, because the shops are of different sizes, large shops have been manually divided so that all shops or shop segments are approximately the same size.

9.4.1 Indoor Scenario 1: Two-Floor of EE building in Queen Mary Campus



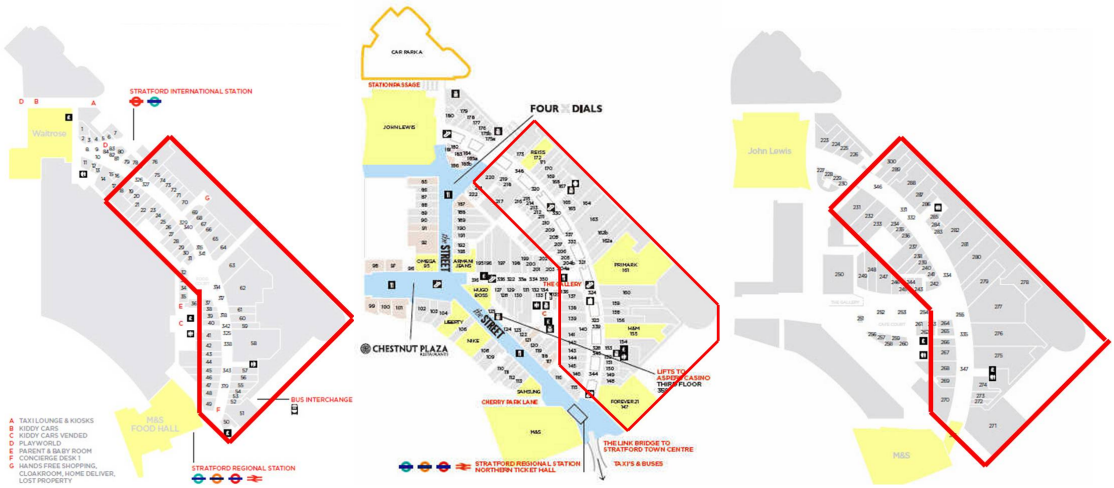
(a) The 2nd floor (b) The 3rd floor

Figure 9.2: The layout of the 2nd and 3rd-floor of EE building in the Queen Mary campus

The 2nd and 3rd-floor of EE building in the Queen Mary campus is used as the first test-bed in this work. In this indoor environment, 500 samples of GSM and WiFi signal strengths have been collected in 15 different rooms on the two floors. There are 21 nearby BSs detected in GSM networks and 20 stable APs found in WiFi networks.

In the experiments, the data are randomly divided into two sets. The first set is treated as training data and is a collection of 400 samples. Their corresponding room location information is known. The other set contains the other 100 samples and is used for room estimation test using only the RSS values, with their correct room number subsequently only used for validation.

9.4.2 Indoor Scenario 2: London Stratford Westfield Shopping Mall



(a) Lower Ground Floor (b) Ground Floor (c) First Floor

Figure 9.3: The layout of the three-floor of London Stratford Westfield shopping mall

In this test-bed, 5323 samples of GSM and WiFi signal strengths are collected in 177 different rooms or room segments on the three floors. There are 41 nearby BSs detected in GSM networks and 199 stable APs found in WiFi networks. Similarly, in the experiments, the data are randomly divided into two equal sets. The first set is treated as the training data set. The other set is used for room estimation test.

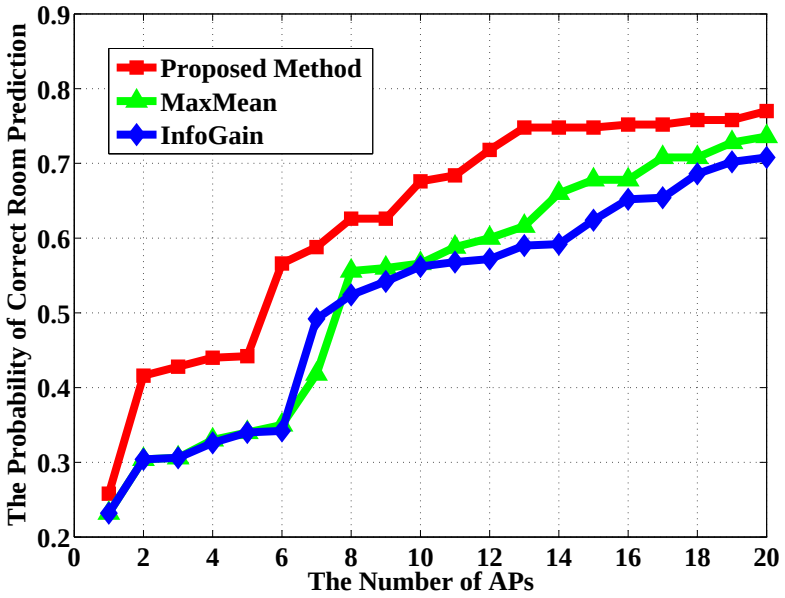


Figure 9.4: The average accuracy of correct room prediction versus the number of APs in indoor scenario 1: EE building in Queen Mary Campus

9.5 Performance Evaluation

9.5.1 The Effects of Transmitters Selection Methods

In the EE office case, to select the best number of APs from the 20 stable APs, the proposed approach (global-PCA) is compared with the MaxMean [31] and InfoGain [52] approaches. (Note that for larger areas, it has been also looked at selecting and rotating to Principal Component axes computed for each cluster. This is sensible as each cluster is different, but each individual cluster is (more) homogeneous. For the InfoGain approach, every room is taken as a grid element. The performance is evaluated in terms of the average accuracy of room estimation, which is defined as the cumulative percentage of estimations within specified errors. Figure 9.4 shows the accuracy comparison between MaxMean, InfoGain and the proposed transmitter selection method. It can be clearly seen that the PCA approach significantly outperforms the traditional methods under the same numbers of the APs. For example, when using 12 APs, the proposed transmitter selection approach reports 71.8% accuracy of room estimation while those of MaxMean

and InfoGain are 60% and 57.2% respectively. Likewise, the proposed transmitter selection approach performs better than the other two methods for the GSM networks. In this comparison, 12 APs and 4 BSs are chosen as the best subset respectively after using PCA. When applying the global PCA into the large area, e.g. the London Stratford Westfield shopping mall, the best number of APs and BSs are 13 and 6 respectively.

However, as mentioned before, the target of this research is to perform location estimation in a large indoor area. Here it might not be a reasonable way to use global PCA to choose a best subset of transmitters relevant to all possible locations. Any one of detectable transmitters cannot be necessarily neglected because each of them takes the important responsibility in the region where it covered. So the global PCA method (a.k.a G-PCA) has been described in chapter 4 section 4.3 is compared with the proposed approach in this chapter where Principal Components are selected within each cluster (a.k.a C-PCA) from the full set of transmitters. For each approach, in each cluster the best transmitters are used, i.e. those that account for most of the variability in the data. Both methods are tested by using different subsets of the RSS in indoor scenario 1 and 2 respectively, as shown in Figure 9.5 and Figure 9.6. These two figures not only show the correct room prediction accuracy, but also illustrate the accuracy of obtaining either the correct room or a neighbouring room or room segment (where a neighbour is defined as the adjacent room segment on the same floor). A marked improvement in accuracy is found using C-PCA, especially for the shopping mall. The reason is that the chosen Principal Components can be quite different in each cluster after a suitable transformation, and C-PCA does not require the selection of a single relevant subset, which G-PCA does. Therefore C-PCA is more scalable.

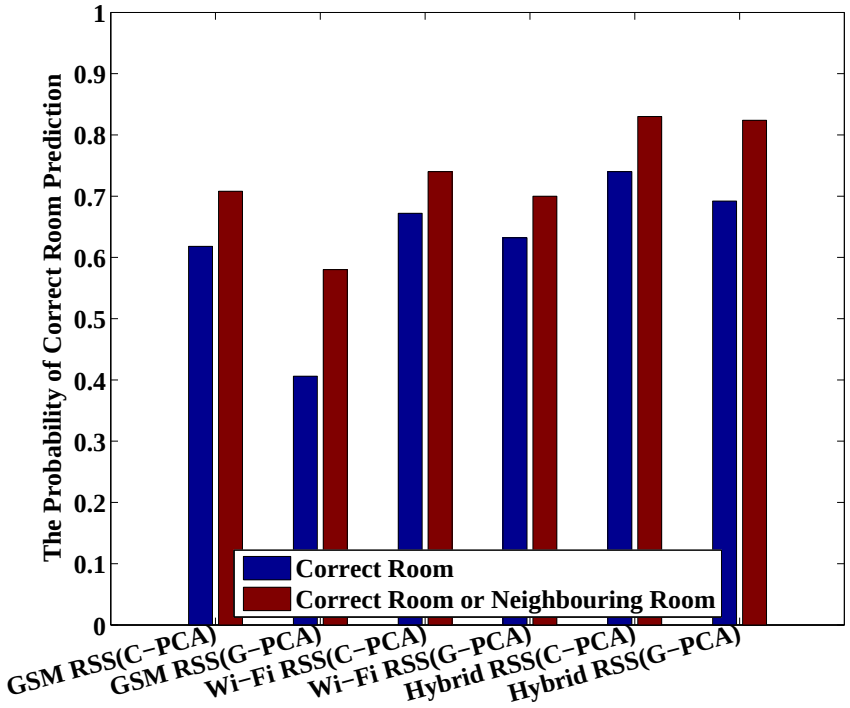


Figure 9.5: The probability of correct room estimation results comparisons between cluster-based PCA and Global PCA methods in three forms of RSS in indoor scenario 1: EE building in Queen Mary Campus.

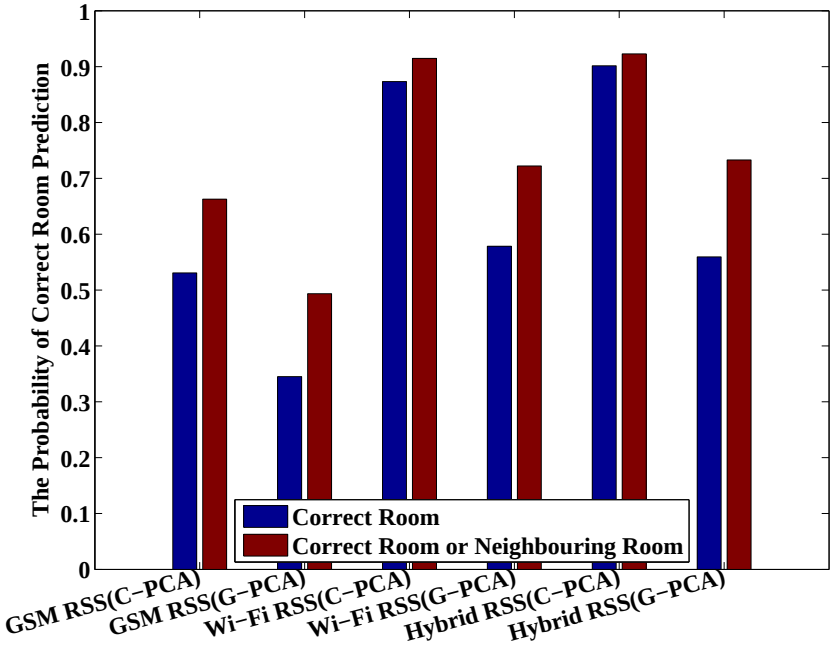
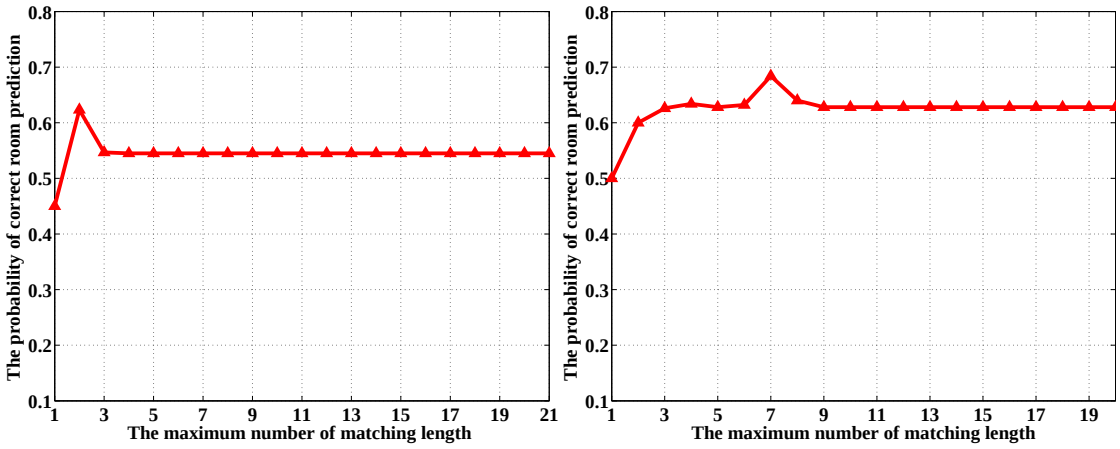


Figure 9.6: The probability of correct room estimation results comparisons between cluster-based PCA and Global PCA methods in three forms of RSS in indoor scenario 2: London Stratford Westfield Shopping Mall

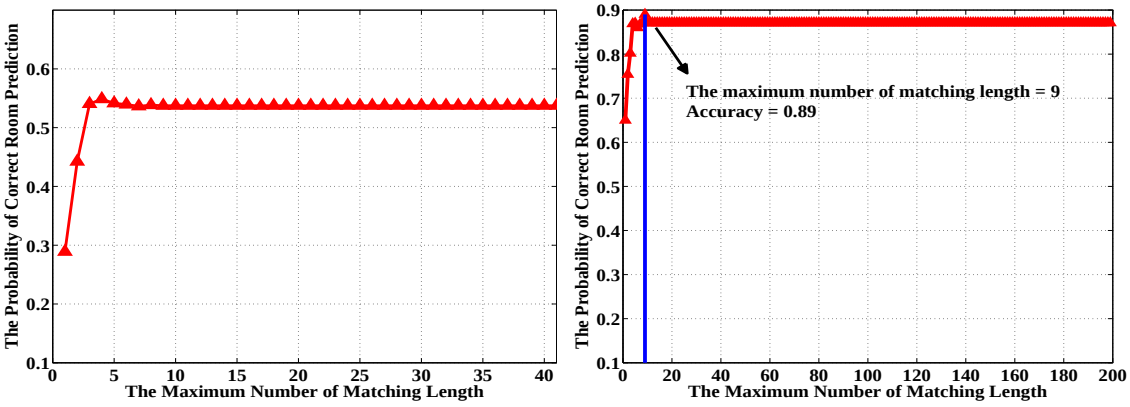
9.5.2 The Effect of the Matching Length in Clustering

The proposed method needs to create different clusters according to ranking patterns of different lengths during the training phase. To balance the trade-off between computational complexity and estimation accuracy, it is important to find the best maximum matching length to create clusters. Figure 9.7 and Figure 9.8 show the room estimation accuracy versus the highest chosen maximum matching length used in clustering in both GSM networks (g) and WiFi networks (w) in these two different scenarios. This corresponds to the depth of the tree constructed during training phase. Taken the EE building in Queen Mary campus for an example, seen from Figure 9.7 (b), it can be found that the estimation accuracy increases as the matching length increases from 1 to 8. However, the predictive accuracy does not monotonically improve along with the increasing matching length. When the maximum allowed matching length is set as 7, inclusion of additional RSS leads to worse rather than better performance. Similarly, for GSM shown in Figure 9.7 (a), the best maximum number length allowed is set as 2. Likewise, for the London Stratford Westfield Shopping Mall, the best maximum number length in GSM networks and WiFi networks are set to 4 and 9 as illustrated in Figure 9.8 (a) and Figure 9.8 (b).



(a) GSM Networks (b) WiFi Networks

Figure 9.7: The probability of correct room prediction versus the maximum number of matching length in clustering in (a) GSM networks and (b) WiFi networks in indoor scenario 1: EE building in Queen Mary Campus



(a) GSM Networks (b) WiFi Networks

Figure 9.8: The probability of correct room prediction versus the maximum matching length in clusters (a) GSM networks and (b) WiFi networks in indoor scenario 2: London Stratford Westfield Shopping Mall

9.5.3 Positioning Performance

The performance of the proposed localisation method is compared with the KNN method in [53] and the KDE method in [72], which assumes RSSs are independent statistically. For the KDE method, the RSS probability density for every room is built. Three forms of RSS are used, viz. GSM RSS, WiFi RSS and both of them (a.k.a hybrid RSS).

Figure 9.9 and Figure 9.10 compare the estimation accuracies of the three different algorithms by using GSM RSS, WiFi RSS and both of them in the two test-beds. For each test-bed, 10 trials for every algorithm are performed and the mean value is plotted, and also the same number of training data points and test data points is used in each trial. From these two figures, it can be clearly seen that the proposed localisation method significantly outperforms the two traditional methods, with a marginal improvement when hybrid RSS data (WiFi RSS and GSM RSS) is used. For instance, for the Stratford Westfield shopping mall, when integrating WiFi RSS with GSM RSS, the proposed method can achieve 93.7% accuracy of correct room prediction, whereas the KNN and KDE methods report 60.1% and 15.3% respectively. Furthermore, it can be observed

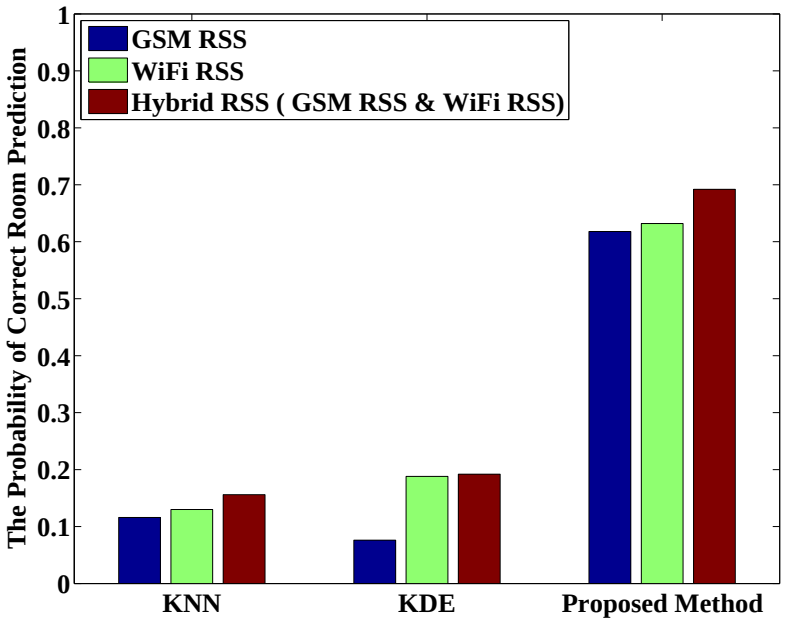


Figure 9.9: The correct room estimation accuracy results for different algorithms in three forms of RSS in indoor Scenario 1: Two-Floor of EE building in Queen Mary Campus

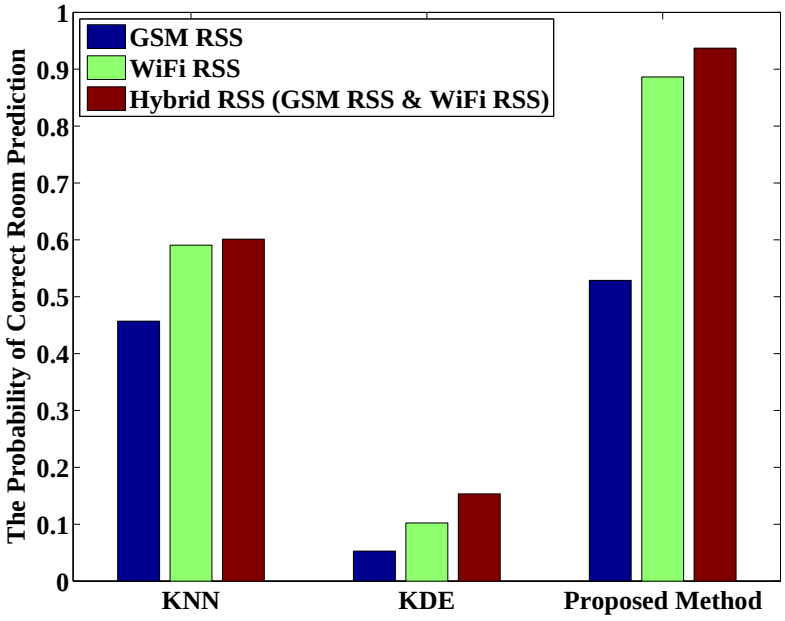


Figure 9.10: The correct room estimation accuracy results for different algorithms in three forms of RSS in indoor Scenario 2: London Stratford Westfield Shopping Mall

that all the three methods based on hybrid RSS data perform better than those based on only GSM signal strength or WiFi signal strength at a small extent. Moreover, it clearly shows, as expected, that using WiFi data can achieve better accuracy than using GSM

data. This is because the variation of GSM signal strengths in different rooms is smaller than that of WiFi signal strengths and there are more APs available. In addition, it can be clearly observed that the performance of the KNN approach used in shopping mall does better than when applied in the EE building in the Queen Mary campus. The difference in the received signal in different rooms in a small-sized environment is relatively smaller than that in a larger-sized environment. In the relatively small area, e.g. EE building in the Queen Mary campus, the signal powers received in any two rooms are very similar. It appears to be more difficult to use the simple "distance" function in signal space to find the nearest match.

9.5.4 Comparison with Kendall's rank correlation

One of the main ideas in the proposed algorithm is to assign a new mobile user to a specific cluster, which is built by the sequence of the transmitter labels sorted by their RSS values in a descending order from training samples. Here the *Kendall tau rank correlation coefficient* [102] is used as a benchmark to evaluate the performance of the proposed method. The Kendall tau rank correlation coefficient is a statistic used to measure the degree of similarity between two sets of ranks given to a same set of objects [102]. This allows a mobile user to be within two or more clusters according to the similarity between any two predefined subsets of relevant transmitters, rather than a specific cluster. Both methods are tested by using different subsets of the RSS in the different indoor scenarios, and the results shown in Figure 9.11 and Figure 9.12.

Seen from these two figures, the proposed method performs only a little better than using Kendall tau rank correlation coefficient (One would expect it to be marginally worse) but is much less time-consuming. For example, for the Stratford Westfield shopping mall data, the average processing time to predict the room ID for one new mobile user with Kendall's tau takes 37.9 seconds, while the proposed method only takes 0.5 seconds (on a small laptop).

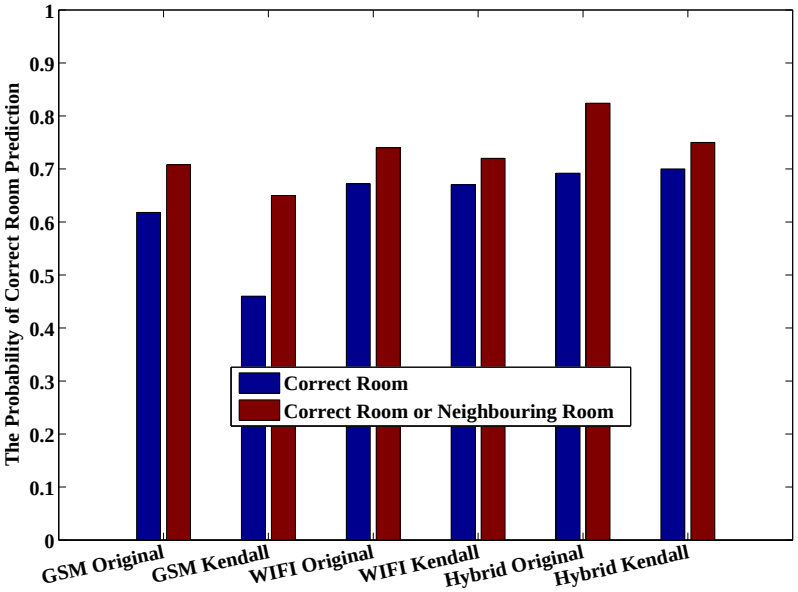


Figure 9.11: The correct room or neighbouring room estimation accuracy results comparisons between the proposed method and the method using Kendall tau rank correlation coefficient in indoor Scenario 1: Two-Floor of EE building in Queen Mary Campus

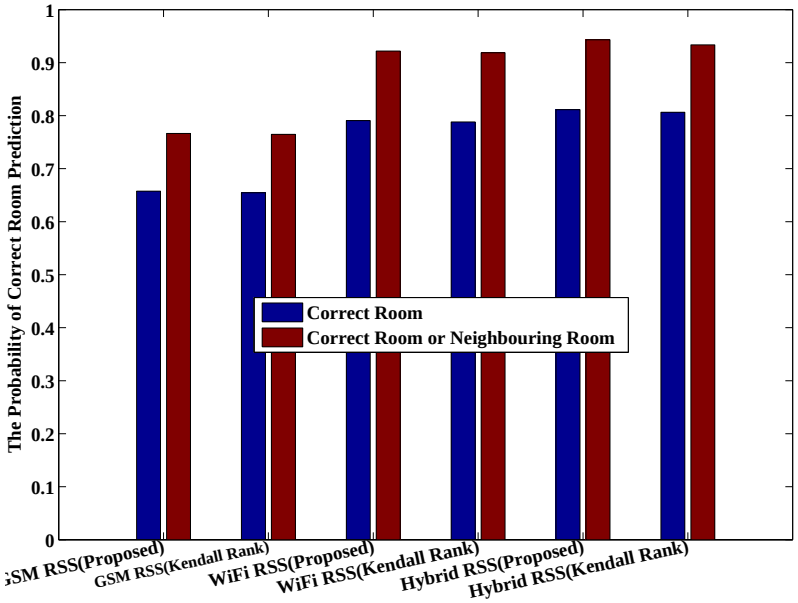


Figure 9.12: The correct room or neighbouring room estimation accuracy results comparisons between the proposed method and the method using Kendall tau rank correlation coefficient in indoor Scenario 2: London Stratford Westfield Shopping Mall

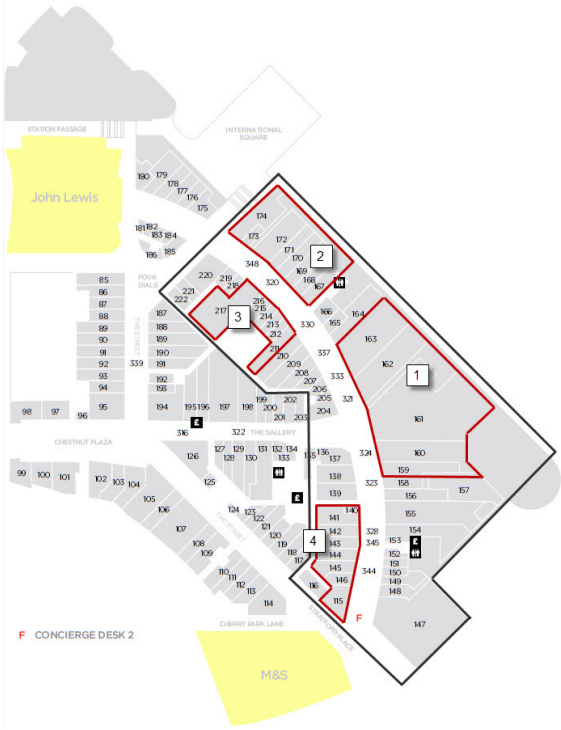


Figure 9.13: The layout of the emergency areas on the ground floor in the London Stratford Westfield Shopping mall

9.5.5 Estimation Accuracy in an Emergency Situation

In this section, 4 different small-size areas are considered as test-beds marked in the Figure 9.13. There are 1613 RSS samples collected on the ground floor. The data points are randomly divided into two equal sets. The first set is considered as the training data set, in which all the APs work well. The other set is treated as the test data set that is further divided into another two sets: one is the secondary training data set and is used for predicting the number of the failed APs and their corresponding IDs around the emergency areas, and the other set is to validate the proposed approach.

For each test-bed, the strongest 5 or 10 APs in that area are chosen to shut down to evaluate the performance of the proposed approach (this will have a considerable effect on the performance of the location estimation procedure.) In the experiments, different scenarios are taken into account: (1) all the APs in the area are in good condition; (2) when the APs fail, but no modification to the procedure is made; (3) using different

numbers of the secondary training data points, e.g. all the test data points, a half of the test data points and a quarter of the test data points, to improve the accuracy under the emergency situation. For each scenario, 5 trials for the proposed scenario are generated and the mean value is plotted. The same number of training data and test data is used in each trial.

The experiment results in different test-beds are described in the following sections: from section 9.5.5.1 to section 9.5.5.4. For example, for test-bed 1, Table 9.1 illustrates the estimation results of the prediction the number of the failed APs and their corresponding IDs. The data in this table shows that the proposed approach can help to find out the failed AP IDs to some extent when there are different numbers of APs are unavailable. Furthermore, as indicated in Figure 9.14, the comparison results in the two sub-figures clearly show that the data update is effective. More specifically, as seen from Figure 9.14 (b), the percentage of the proportion of times that the correct room obtained in the proposed approach (based on the adapted RSS data) according to all the test data, half the test data and a quarter of the test data is 86.6%, 80.5% and 77.5% respectively; whereas without calibrating the radio map is e.g. 62.2% when there are 10 APs are unavailable in test-bed 1. Similarly, it can be observed from Figure 9.14 (a) that the proposed methods with calibration can perform better than that without applying any correction for the changes under the condition that 5 APs are failed.

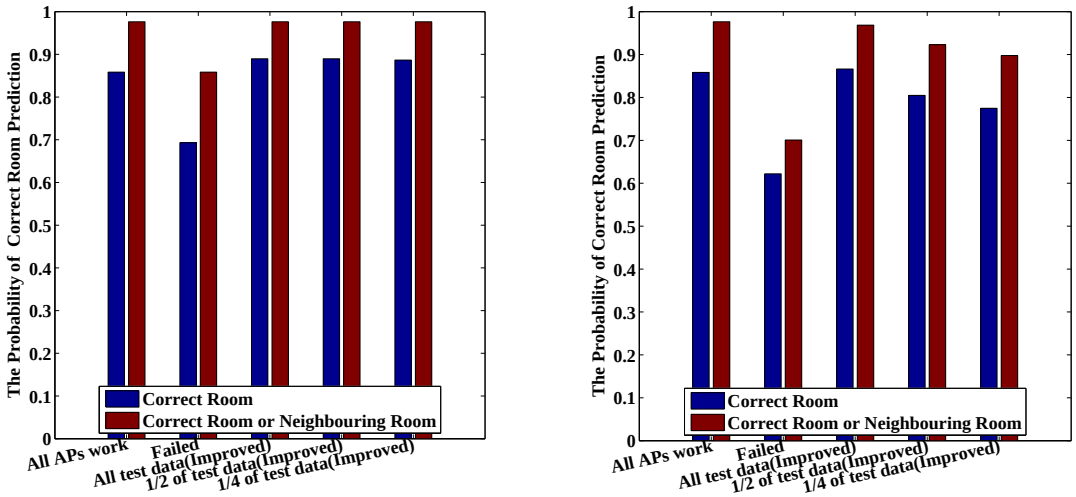
According to Algorithm (9.3), the average number of APs detected by every training data point in the whole area is needed to be calculated at first. But for a large area, e.g. the ground floor of the Westfield shopping mall, it is difficult to make sure that the average number of APs covered by each store or store segment is the same value. The distribution of APs is not uniform in the real environment. That is why the accuracy results of estimated failed APs in the other three test-beds (test-bed 2, 3 and 4) are slightly worse than the result in test-bed 1. But it still can be seen that the proposed method make improvements in location estimation in emergency situation as shown in Figure 9.15, 9.16 and 9.17. In addition, it can be conclude that using the small number

of sample data in an emergence environment can support a good accuracy to figure out the failed AP number and corresponding IDs.

9.5.5.1 Test-bed 1: Room 159-Room 163

Table 9.1: The probability of estimating the failed 5 and 10 APs IDs in Test-bed 1

	All the test data points	**1/2 of the test data points**					**1/4 of the test data points**				
$N_{q_{failed}}$	5	5	5	5	5	5	5	5	4	5	5
Probability	1	1	1	1	1	1	1	1	0.8	1	1
$N_{q_{failed}}$	7	7	7	6	7	7	6	7	7	6	7
Probability	0.7	0.7	0.7	0.6	0.7	0.7	0.6	0.7	0.7	0.6	0.7



(a) 5 APs are failed

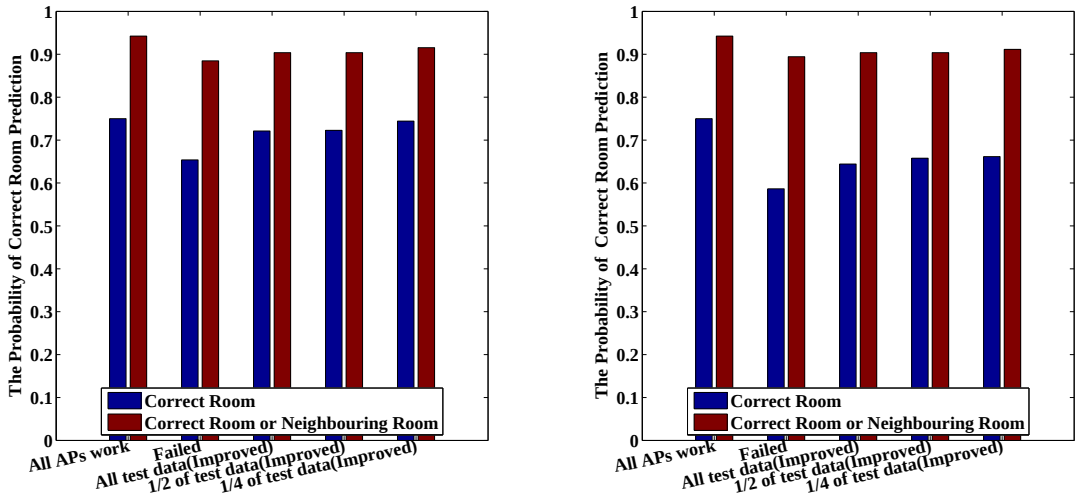
(b) 10 APs are failed

Figure 9.14: The comparison between the approach when all the APs work well, without taking any measurements under emergency situation, and using all the test data points, half of the test data points and a quarter of the test data points to make improvements in emergency situation in Test-bed 1: (a) 5 APs are failed and (b) 10 APs are failed.

9.5.5.2 Test-bed 2: Room 167-Room 174

Table 9.2: The probability of estimating the failed 5 and 10 APs IDs in Test-bed 2

	All the test data points	**1/2 of the test data points**					**1/4 of the test data points**				
$N_{q_{failed}}$	4	5	4	4	4	4	6	5	5	5	4
Probability	0.8	1	0.8	0.8	0.8	0.8	1	1	0.8	1	0.8
$N_{q_{failed}}$	6	6	6	6	7	6	7	6	7	6	8
Probability	0.6	0.6	0.6	0.6	0.7	0.6	0.6	0.6	0.7	0.6	.6 0.7



(a) 5 APs are failed

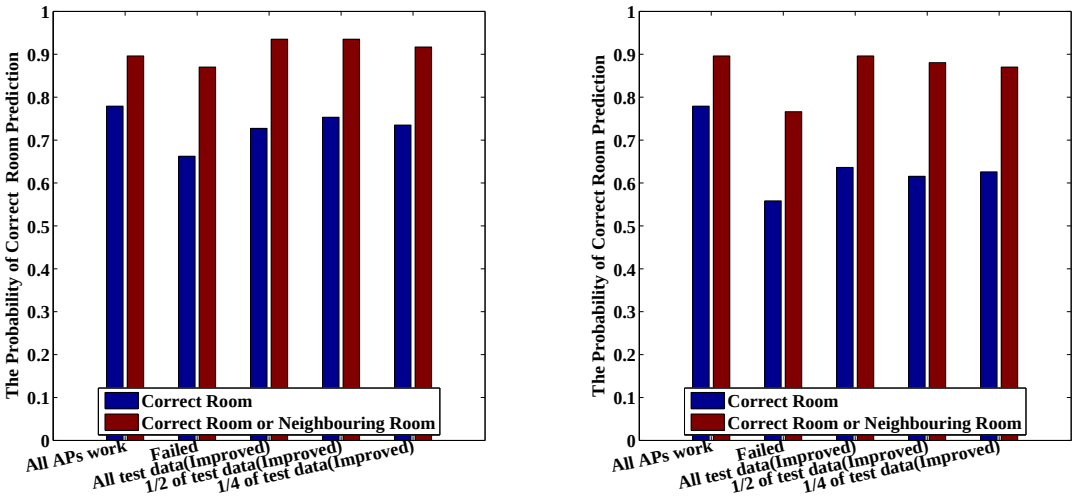
(b) 10 APs are failed

Figure 9.15: The comparison between the approach when all the APs work well, without taking any measurements under emergency situation, and using all the test data points, half of the test data points and a quarter of the test data points to make improvements in emergency situation in Test-bed 2: (a) 5 APs are failed and (b) 10 APs are failed.

9.5.5.3 Test-bed 3: Room 147-Room 154

Table 9.3: The probability of estimating the failed 5 and 10 APs IDs in Test-bed 3

	All the test data points	**1/2 of the test data points**					**1/4 of the test data points**				
$N_{q_{failed}}$	5	5	5	6	4	5	5	5	6	4	4
Probability	1	1	1	1	0.8	1	1	1	1	0.8	0.8
$N_{q_{failed}}$	7	6	7	7	7	7	0.7	7	6	7	5
Probability	0.7	0.6	0.7	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.5



(a) 5 APs are failed

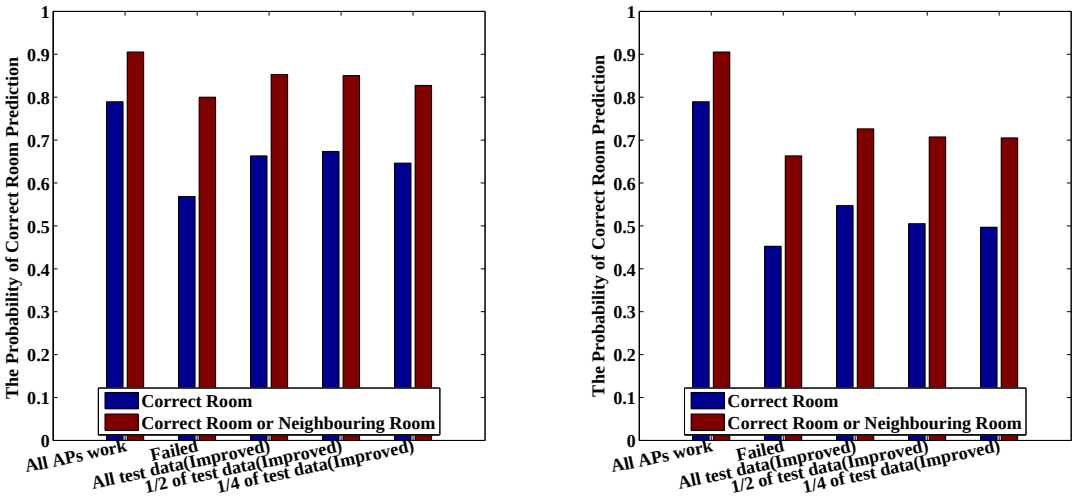
(b) 10 APs are failed

Figure 9.16: The comparison between the approach when all the APs work well, without taking any measurements under emergency situation, and using all the test data points, half of the test data points and a quarter of the test data points to make improvements in emergency situation in Test-bed 3: (a) 5 APs are failed and (b) 10 APs are failed.

9.5.5.4 Test-bed 4: Room 141-Room 146 and Room 115

Table 9.4: The probability of estimating the failed 5 and 10 APs IDs in Test-bed 4

	All the test data points	**1/2 of the test data points**					**1/4 of the test data points**				
$N_{q_{failed}}$	4	4	3	3	4	3	3	3	3	3	2
Probability	0.8	0.6	0.6	0.6	0.8	0.6	0.6	0.6	0.6	0.6	0.4
$N_{q_{failed}}$	6	6	5	6	5	6	6	7	6	6	6
Probability	0.6	0.6	0.5	0.6	0.5	0.6	0.6	0.7	0.6	0.4	0.6



(a) 5 APs are failed

(b) 10 APs are failed

Figure 9.17: The comparison between the approach when all the APs work well, without taking any measurements under emergency situation, and using all the test data points, half of the test data points and a quarter of the test data points to make improvements in emergency situation in Test-bed 4: (a) 5 APs are failed and (b) 10 APs are failed.

9.6 Summary

This chapter has described solutions to estimation location in terms of room number (or room segment) or neighbouring room number in a large scale multi-story indoor environment. First, a novel hybrid RSS-based room estimation approach in static environment has been proposed. To evaluate the performance of the proposed cluster-based deterministic algorithm, real RSS from both WiFi and GSM networks have been collected on the two floors of EE building in the Queen Mary campus and the three floors of Stratford Westfield shopping mall in London. The results indicate using the hybrid RSS can improve the estimation accuracy in a multi-story building compared with the traditional algorithms. Secondly, how to locate mobile users in situations where contiguous APs fail at a Shopping Mall has been considered. An improved approach to allow WiFi radio maps to be adapted to the emergency situation has also been introduced. The improvements in room estimation accuracy in simple emergency scenarios are tested and the results show the GSM data can be used as a backup to ameliorate the loss of accuracy when APs fail. In failure scenarios at the mall, when 5 contiguous APs are chosen to fail, an accuracy increase of 25% in the room segment prediction is found. The comparison is made between the observed WiFi RSS and calibrated WiFi RSS using the GSM RSS correction. The approach leverages itself using GSM RSS data to find out the failed APs in a small-sized area.

Chapter 10

Conclusions and Future Work

10.1 Conclusion

The focus of this thesis is the problem of location estimation and prediction of coverage patterns. Such techniques can be used to support a wide range of applications and support SON functions, e.g. in LTE.

Location estimation based on RSS is problematic as the accuracy is inevitably limited. It is affected by weather, chip set, antenna, phone orientation, etc. The overall aim of this work is to develop mechanisms to enhance the accuracy of location estimation based on RSS, which is readily available from GSM base stations and Wi-Fi access points. RSS is attractive as it can be collected from COTS smartphones and does not require additional equipment (though accuracy can be improved by the use of additional sensors). When a person wants to know his or her current location, the mobile client installed in his or her mobile phone can collect the raw RSS data from each of the location sources, and then send these data to a location server. The location server generates the location from the new data by using advanced hybrid positioning algorithms based on the database. By leveraging the strengths of more than one underlying position technology, the proposed location system can provide the better possible location available in any environment.

In this thesis, the proposed localisation mechanism can run on a real-time operation. For example, the proposed algorithms and the radio map can be located in a server. The algorithms are at the back end of an HTTP server. A servlet can invoke the specialised MATLAB code. Therefore, the client app can send the RSS data to the server and the server sends back the location estimate.

After introducing and motivating the problem in chapter 1, a review of fundamental concepts and relevant literature on location estimation and coverage prediction was provided in chapter 2. A detailed introduction of location fingerprinting techniques was given in chapter 3. A run-time positioning measurement mechanism for outdoor environment was proposed in chapter 4 to provide a better precision of user positioning using RSS. Additionally, a transmitter selection method and a clustering scheme were also introduced to better partition the wireless environment into different homogenous regions based on users' RSS feedback. To improve outdoor positioning accuracy, two different novel localisation methods were proposed respectively in chapter 5 and chapter 6, and simulation and experimental results shows they both can provide higher accuracy for outdoor user positioning. To improve the performances of these two methods in different environmental conditions, chapter 7 introduced a mechanism that allows radio maps to be adapted to environmental changes. A filtering model was briefly introduced in chapter 8 to adapt to a dynamic environment and a nonparametric probability algorithm was developed to build radio coverage models in each area created by the individual clusters. In chapter 9, an approach to large scale indoor location estimation was described. This was then extended to a mechanism that integrates RSS data from both GSM and WiFi networks by using clustering and rank order matching.

An advantage of the approaches described in this thesis is that none require the cooperation of a network operator.

Five of the most significant contributions of this thesis are discussed below.

- **Validation of estimation of locations based on RSS distribution models by using real data sets collected from outside and inside**

  Outside, four different scenarios have been considered in order to compare the outdoor localisation approaches proposed in this thesis. These scenarios can be divided into two: those that depend on simulated data and those based on real data. For the real data, two scenarios (Queen Mary campus and a music festival in London Victoria Park) are considered. As both of them are essentially outdoors the validation used GPS. Two simulated scenarios were also considered in outdoor environment, a regular grid with different shadowing deviations in each grid element, and data collected by radio models on the island of Jersey. For inside buildings, two different sizes of multi-floor real test-beds (viz. two-floors of EE building in the Queen Mary campus and three-floors of London Stratford Westfield shopping mall) are used to perform to validate the proposed indoor localisation approaches.

- **A clustering scheme for outdoor localisation**

  In the proposed run-time positioning measurement mechanisms, the proposed clustering scheme plays an important role. This influences the precision of position accuracy and coverage prediction and allows models of coverage to be created for individual clusters. It is shown that the partitioning into clusters outperforms grid-based and global-based methods in some of the specific scenarios. Mobile location is estimated by the cluster it belongs to and its relative location in that cluster and coverage model is also built in that cluster to estimate radio coverage probability. The clusters are created by analysis of RSS data points collected from training data and further improved by users of the applications during operation.

  The novel features of the clustering scheme are: a) the use of deviations from the observed path loss model for each RSS component rather than the raw RSS. This also results in the clusters being invariant to the BS/RS power; b) the accurate estimation of the cluster membership probability and the number of clusters to manage the trade-off between cluster size and accuracy of cluster modelling using

the VPM; c) it allows selective additional data collection to enhance accuracy, e.g. near cluster boundaries; d) it is not sensitive to transmit power changes; e) it can be coupled with location tracking and other non RSS data to improve accuracy.

- **Improved outdoor localisation approaches for static and dynamic environments**

  The use of PCA is shown to offer an efficient mechanism to utilize information from all detectable transmitters and to retain correlations in the RSS by rotating to orthogonal dimensions in each cluster. The PCs are generated through a transformation relevant to each cluster in the RSS data, and the selected BSs and the data reduction can be different in each cluster.

  State-of-the-art deterministic and probabilistic location estimation approaches were proposed.

  In the deterministic framework, an improved PCA-Intersection method to outdoor fingerprinting location estimation based on clustering RSS from BSs was tested in four different scenarios (rural, urban, and suburban). Results presented show that the proposed scheme finds more accurate locations and outperforms the traditional probabilistic approach and KNN approach in the experiments.

  In the probabilistic framework, considering the high correlation between signal strengths from different transmitters, PCA-KDE was utilised to maintain the most important RSS characteristic information and reducing useless signal information. The KDE was then used for location estimation. This nonparametric approach provided a powerful tool set for modelling of spatio-temporal RSS properties based on the training-based fingerprinting approach. KDE and relevant parameter choices were studied both theoretically and experimentally. It was shown that the PCA-KDE estimate is superior to the original KDE used in both simulated and real data in terms of positioning accuracy. In the orthogonal space the joint probabilities are computed rapidly by simple multiplication.

  Moreover, a mechanism has been introduced to allow radio maps to be adapted

to environmental changes. Only one full radio map for a specific environment is needed. Based on this, an updating pattern for a new environment can be created. The calibrated RSS data can be regarded as measured in the same environment as the reference training data. The improvement in location estimation accuracy is tested, and the results show that the proposed algorithms achieve a considerable advantage over previous static fingerprint-based techniques in the test-beds.

- **A nonparametric probability approach for modelling radio coverage**

  The method proposed in this thesis mitigates inaccuracies resulting from changes in the physical environment as it provides a way of detecting them. Firstly, the self-training semi-supervised mechanism was applied to the proposed mechanism which removes the need to know the precise location of most of the recorded fingerprints during the training process. A few carefully selected key points represent regions of fingerprints. Secondly, to provide reliable radio coverage in the wireless environment, KDE was utilised again to build an accurate coverage model in each cluster based on the clustering scheme. A filter model was proposed to detect significant changes in the radio map.

- **Hybrid RSS-based Room Estimation Method**

  A novel hybrid RSS-based room estimation approach by a hierarchical partitioning scheme in multi-story indoor environment was described and validated in the real environment. The results indicate using the hybrid RSS can improve the estimation accuracy in multi-story building compared with the traditional algorithms. Additionally, a method used GSM data to figure out the failed APs in a simple emergency scenario, and it demonstrated to allow improved accuracy when APs fail, e.g. in emergency situations.

10.2 Possible Extensions and limitations

Several technical issues or aspects remain to be explored before continuing to this work:

- The main limitation of the proposed system is the sensitivity of the training location fingerprints to environmental changes and device characteristics. Furthermore, the laborious nature of fingerprint collection hinders the scalability of the proposed technique to large environments. An interesting solution for overcoming these limitations is the use of dynamically built radio maps based on real-time sensing of the environment. This is being investigated.

- With the author's understanding, location-based systems that depend on RSS alone have been criticised because of inaccuracies, due to changes in humidity, temperature, and physical environment. Methods that use auxiliary active RFID tags have been proposed for high indoor accuracy, but this is not ideal for general use in larger areas. As mentioned in chapter 8, the proposed method mitigates inaccuracies resulting from changes in the physical environment because RSS coverage models are created and can be used to check for discrepancies over time and the model of deviations is corrected across a threshold and also accommodates changes in power at the transmitter in chapter 4. However, as the experimental results in the Music Festival in London Victoria Park show (chapter 7), the influence of environmental factors, such as temperature, humidity and rain amount, on the accuracy of positioning will be concentrated on. For example, for humidity the author can collect outdoor data in rain and sunshine and cumulative data collection can build up a suitable data base for different conditions. The further work aims to collect systematic data to validate the proposed calibration method in changeable environment so that appropriate path loss models can be constructed.

- The limitations of the indoor research in chapter 9 are the difficulty for obtaining exactly x-axis and y-axis information in the indoor environment and simplistic emergency scenarios are used. Extension of the work will focus on how to record the geo-location or relative location in the Westfield shopping mall by using a smart phone app and integrating auxiliary geographical and smart phone sensor information to improve estimation accuracy (such as magnetic field). Moreover,

the different degrees of damage and the associated communication strategy will be also considered.

- Further work intends to improve the estimation method when appropriate by predicting future locations based on previous locations (e.g. using Kalman filtering). This temporal averaging will be used to reduce battery usage, as if the context indicates that prediction will be accurate, and then sampling RSS will be unnecessary. It also provides a way of smoothing the data akin to taking repeated measurements. The author plans to use auxiliary topographic knowledge so as to decide when and when not to use the predictor. It is more important to recognise when filters could be useful than to have to elaborate prediction techniques. For example at traffic lights, a linear predictor, such as the Kalman filter, is often not sensible. It is better not to predict at all. Additionally, investigation into the selective use of use of accelerometers in the smart phones, again based on auxiliary knowledge, to support identification of location, such as lift or bus or escalator, which can have discriminatory profiles and act as landmarks will be taken into account.

Appendix A. Acknowledgements

A special thanks to Aircom International Ltd for providing their network planning tool ASSET 3G and for the data provided on pilot signal strengths for the island of Jersey, and Loud Sound Company for offering the map for Grid Field Day for the Music Festival in London Victoria Park.

Appendix B. Author's Publications

Journal Papers

1. K. Li, P. Jiang, E. L. Bodanese and J. Bigham, "Outdoor Location Estimation Using Received Signal Strength Feedback," IEEE Communications Letters, vol. 16, no. 7, pp. 978-981, July 2012.

2. K. Li, J. Bigham and L. Tokarchuk, "Validation of a Probabilistic Approach to Outdoor Localization", IEEE Wireless Communications Letters, vol. 2, no. 2, pp. 167-170, May 2013.

3. K. Li, J. Bigham, E. L. Bodanese and L. Tokarchuk "Outdoor Location Estimation in Changeable Environments", IEEE Communications Letters, Accepted for publication.

Conference Papers

1. K. Li, P. Jiang and J. Bigham, "Partitioning the Wireless Environment for Determining Radio Coverage and Traffic Distribution with User Feedback", in *Proceedings of the 17th National Conference on Communications*, Bangalore, India, Jan. 2010.

2. K. Li, P. Jiang and J. Bigham, "Cluster and grid based weighted K-Nearest Neighours for outdoor location estimation", in *Proceedings of IET International Conference on Communication Technology and Application*, Beijing, China, pp. 833-838, Oct. 2011.

3. K. Li, P. Jiang, E. L. Bodanese and J. Bigham, "Real Time Radio Coverage Monitoring in Self-organizing Networks with User Feedback", in *Proceedings of 5th International Workshop on Selected Topics in Mobile and Wireless Computing*, Barcelona, Spain, Oct. 2012.

4. K. Li, J. Bigham, E. L. Bodanese and L. Tokarchuk, "Location Estimation in Large Indoor Multi-floor Buildings using Hybrid Networks", in *Proceedings of IEEE Wireless Commun. and Networking Conf.(WCNC)*, Shanghai, China, April 2013.

5. K. Li, J. Bigham, L. Tokarchuk and E. L. Bodanese, "A Probabilistic Approach to Outdoor Localization Using The Clustering and Principal Component Transformations", in *Proceedings of the 9th International Wireless Communications & Mobile Computing Conf. (IWCMC)*, Cagliari, Sardinia, Italy, July 2013.

References

[1] P. Jiang. Cooperative control of relay based cellular networks. In *A thesis of Queen Mary University of London for the degree of Doctor of Philosophy*, Electronic Engineering and Computer Science, Queen Mary, University of London, 2009. xiii, 1, 12

[2] A. R. Mishra. *Fundamentals of Cellular Network Planning and Optimisation: 2G/2.5G/3G...Evolution to 4G.* Wiley. com, 2004. xiii, 12, 13

[3] D. W. Scott. *Multivariate Density Estimation.* John Wiley & Sons, 1992. xviii, 99, 100, 101

[4] 3GPP TR 36.902. Self-configuring and self-optimizing network (SON) use cases and solutions. June 2010. 1

[5] J. L. Van den Berg, R. Litjens, A. Eisenblätter, M. Amirijoo, O. Linnell, C. Blondia, T. Kürner, N. Scully, J. Oszmianski, and L. C. Schmelz. Self-organisation in future mobile communication networks. In *Proceedings of ICT Mobile Summit 2008*, Stockholm, Sweden, 2008. 1

[6] P. Jiang, J. Bigham, and M. Anas Khan. Distributed algorithm for real time cooperative synthesis of wireless cell coverage patterns. *IEEE Communications Letters*, 12(9):702–704, 2008. 5, 29

[7] P. Jiang, J. Bigham, R. Dubrovka, and J. Wu. A statistical radio coverage prediction approach for cooperative control in relay based cellular networks. In *Proceed-*

ings of the Fourth European Conference on Antennas and Propagation (EuCAP), Barcelona, Spain, April 2010. 5, 29

[8] S. A. Zekavat and R. M. Buehrer. *Handbook of Position Location: Theory, Practice and Advances*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2012. 7, 26, 27, 28

[9] A. Goldsmith. *Wireless communications*. Cambridge Univ Pr, 2005. 13

[10] T. K. Sarkar, Z. Ji, K. Kim, A. Medouri, and M. Salazar-Palma. A survey of various propagation models for mobile communication. *IEEE Antennas and Propagation Magazine*, 45(3):51–82, 2003. 14, 23, 126

[11] T. S. Rappaport. *Wireless communications: principles and practice*. Prentice Hall, 2002. 14, 62

[12] G. Giorgetti, S. K. S. Gupta, and G. Manes. Localization using signal strength: to range or not to range. In *Proceedings of the first ACM international workshop on Mobile entity localization and tracking in GPS-less environments*, pages 91–96, September 2008. 15

[13] B. J. Dil and P. J. M. Havinga. Rss-based localization with different antenna orientations. In *Proceedings of Australasian Telecommunication Networks and Applications Conference (ATNAC)*, 2010. 15

[14] H. Linde. On aspects of indoor localization. In *A thesis of University of Dortmund for the degree of Doctor of Philosophy*, University of Dortmund, August 2006. 17

[15] M. I. Silventoinen and T. Rantalainen. Mobile station emergency locating in gsm. In *Proceedings of IEEE International Conference on Personal Wireless Communications*, pages 232–238, 1996. 17

[16] N. J. Thomas. Techniques for mobile location estimation in umts. In *A thesis of University of Edinburgh for the degree of Doctor of Philosophy*, College of Science and Engineering, School of Engineering and Electronics, University of Edinburgh, December 2001. 17

[17] W. H. Foy. Position-location solutions by taylor-series estimation. *IEEE Transactions on Aerospace and Electronic Systems*, AES-12(2):187–194, March 1976. 17

[18] X. Wang, Z. Wang, and B. O'Dea. A toa-based location algorithm reducing the errors due to non-line-of-sight (NLOS) propagation. *IEEE Transactions on Vehicular Technology*, 52(1):112–116, 2003. 18

[19] J. J. Caffery. *Wireless location in CDMA cellular radio systems*. Kluwer Academic, 2000. 18, 19, 22

[20] P. H. Tseng, C. L. Chen, and K. T. Feng. An unified kalman tracking technique for wireless location systems. In *Proceedings of IEEE 2nd International Symposium on Wireless Pervasive Computing (ISWPC)*, pages 1001–1011, San Juan, Puerto Rico, February 2007. 18

[21] Y. T. Chan and K. C. Ho. A simple and efficient estimator for hyperbolic location. *IEEE Transactions on Signal Processing*, 42(8):1905–1915, 1994. 19

[22] M. Najar and J. Vidal. Kalman tracking based on tdoa for umts mobile location. In *Proceedings of IEEE Int' l Symp. Personal, Indoor and Mobile Radio Comm.*, pages 45–49, September 2001. 19

[23] B. L. Le, K. Ahmed, and H. Tsuji. Mobile location estimator with nlos mitigation using kalman filtering. In *Proceedings of IEEE Conf. Wireless Comm. and Networking*, pages 1969–1973, March 2003. 19

[24] Y. Zhao. Standardization of mobile phone positioning for 3g systems. *IEEE Communications Magazine*, 40(7):108–116, 2002. 20

[25] A. H. Sayed, A. Tarighat, and N. Khajehnouri. Network-based wireless location: challenges faced in developing techniques for accurate wireless location information. *IEEE Signal Processing Magazine*, 22(4):24–40, 2005. 20

[26] K. J. Krizman, T. E. Biedka, and T. S. Rappaport. Wireless position location: fundamentals, implementation strategies, and sources of error. In *Proceedings of the 47th IEEE Vehicular Technology Conference (VTC)*, pages 919–923, May 1997. 20

[27] J. Kennedy and M. C. Sullivan. Direction finding and smart antennas using software radio architectures. *IEEE Commun. Magazine*, pages 62–68, May 1995. 21

[28] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986. 21

[29] N. Czink, E. Bonek, X. Yin, and B. Fleury. Cluster angular spreads in a mimo indoor propagation environment. In *Proceedings of the 16th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pages 664–668, September 2005. 21

[30] A. Krishnakumar and P. Krishnan. On the accuracy of signal strength-based estimation techniques. In *Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies (Infocom)*, pages 642–650, March 2005. 23

[31] M. A. Youssef, A. Agrawala, and A. Udaya Shankar. Wlan location determination via clustering and probability distributions. In *Proceedings of the 1st IEEE Int. Conf. on Pervasive Computing and Commun.*, pages 143–150, March 2003. 23, 34, 37, 38, 42, 53, 54, 140, 151

[32] J. Hightower and G. Borriello. Particle filters for location estimation in ubiquitous computing: A case study. In *Proceedings of International Conference on Ubiquitous Computing (UbiComp)*, pages 88–106, 2004. 23

[33] C. H. Chang and W. Liao. A probabilistic model for relative location estimation in wireless sensor networks. *IEEE Communications Letter*, 13:893–895, 2009. 23, 42

[34] R. Peng and M. L. Sichitiu. Probabilistic localization for outdoor wireless sensor networks. In *Proceedings of ACM Mobile Computing and Communications (SIGMOBILE)*, pages 53–64, January 2007. 24

[35] N.Patwari. Location estimation in sensor networks. In *A thesis of University of Michigan for the degree of Doctor of Philosophy*, University of Michigan, 2005. 24

[36] N. Patwari, R. J. O'Dea, and Y.Wang. Relative location in wireless networks. In *Proceedings of the IEEE Vehicular Technology Conference (VTC)*, May 2001. 24

[37] T. He, C.Huang, B. M. Blum, J. Stankovic, and T. Abdelzaher. Range-free localizationschemes for large scale sensor networks. In *Proceedings of the 9th annual international conference on Mobile computing and networking (MobiCom)*, pages 81–95, September 2003. 24

[38] C. Liu, K. Wu, and T. He. Sensor localization with ring overlapping based on comparison of received signal strength indicator. In *Proceedings of IEEE International Conference onIEEE Mobile Ad-hoc and Sensor Systems (MASS)*, pages 516–518, October 2004. 24

[39] K. Yedavalli, B. Krishnamachari, S. Ravula, and B. Srinivasan. Ecolocation: A sequence based technique for rf-only localization in wireless sensor networks. In *Proceedings of the 4th Int. Symp. on Inform. Processing in Sensor Networks (IPSN)*, pages 285–292, April 2005. 24

[40] S. Venkatraman and J. Caffery. Hybrid toa/aoa techniques for mobile location in non-line-of-sight environments. In *Proceedings of IEEE Wireless Commun. and Networking Conf.(WCNC)*, pages 274–278, March 2004. 25

[41] L. Cong and W. Zhuang. Hybrid tdoa/aoa mobile user location for wideband cdma cellular systems. *IEEE Transactions on Wireless Communications*, 1(3):439–447, July 2002. 25

[42] N. J. Thomas, D. G. M. Cruickshank, and D.I. Laurenson. Performance of a tdoa-aoa hybrid mobile location system. In *Proceedings of the 2nd International Conference on 3G Mobile Communication Technologies*, pages 216–220, 2001. 25

[43] J. Hightower and G. Borriello. Location systems for ubiquitous computing. *Computer*, 34(8):57–66, 2001. 25

[44] S. Bartels. Wifi location system investigation. In *A thesis of University of Waikato for the degree of Doctor of Philosophy*, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 2005. 26

[45] G. M. Djuknic and R. E. Richton. Geolocation and assisted gps. *IEEE Comput.*, 34(2):123–125, 2001. 27

[46] J. Werb and C. Lanzl. Designing a positioning system for finding things and people indoors. *IEEE Spectr.*, 35(9):71–78, 1998. 27

[47] L. M. Ni, Y. Liu, Y. C. Lau, and A. P. Patil. Landmarc: Indoor location sensing using active rfid. *Wireless Networks*, 10:701–710, 2004. 28, 39, 140

[48] L. Du, J. Biaham, and L. Cuthbert. A bubble oscillation algorithm for distributed geographic load balancing in mobile networks. In *Proceedings of the Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies (Infocom)*, pages 330–338, 2004. 29

[49] N. S. Adawi, H. L. Bertoni, J. R. Child, W. A. Daniel, J. E. Dettra, R. P. Eckert, Jr. E. H. Flath, R. T. Forrest, W. C. Y. Lee, S. R. McConoughey, J. P. Murray, H. Sachs, G. L. Schrenk, N. H. Shepherd, and F. D. Shipley. Coverage prediction for mobile radio systems operating in the 800/900 mhz frequency range. *IEEE Transactions on Vehicular Technology*, 37(1):3–72, February 1988. 30

[50] T. Kurner and A. Meier. Prediction of outdoor and outdoor-to-indoor coverage in urban areas at 1.8 ghz. *IEEE Journal on Selected Area in Communications*, 20(3), April 2002. 30

[51] L. Juan-Llacer, L. Ramos, and N. Cardona. Application of some theoretical models for coverage prediction in macrocell urban environments. *IEEE Transactions on Vehicular Technology*, 48(5), September 1999. 30

[52] J. Yin, Y. Chen, X. Chai, and Q. Yang. Power-efficient access-point selection for indoor location estimation. *IEEE Transactions on Knowledge and Data Engineering*, 18(7):877–888, July 2006. 34, 37, 38, 44, 53, 151

[53] P. Bahl and V. N. Padmanabhan. Radar: An in-building rf-based user location and tracking system. In *Proceedings of IEEE 19th Annu. Joint Conf. of the IEEE Comput. and Commun. Soc.*, pages 775–784, March 2000. 35, 39, 40, 44, 140, 155

[54] K. Kaemarungsi and P. Krishnamurthy. Modeling of indoor positioning systems based on location fingerprinting. In *Proceedings of the 23rd of IEEE 19th Annu. Joint Conf. of the IEEE Comput. and Commun. Soc.*, pages 1012–1022, March 2004. 35, 37

[55] K. Kaemarungsi and P. Krishnamurthy. Properties of indoor received signal strength for wlan location fingerprinting. In *Proceedings of the 1st International Conference on Mobile and Ubiquitous Systems: Networking and Services (MOBIQUITOUS)*, pages 14–23, 2004. 35

[56] H. Leppkoski, S. Tikkinen, and J. Takala. Optimizing radio map for wlan fingerprinting. In *Proceedings of the Ubiquitous Positioning Indoor Navigation and Location Based Service (UPINLBS)*, October 2010. 35

[57] M. Bshara and L. V. Biesen. Localization in wimax networks depending on the available rss-based measurements. *International Journal on Advances in Systems and Measurements*, 2:214–223, 2009. 36

[58] A. Haeberlen and A. Rudys. Practical robust localization over large-scale 802.11 wireless networks. In *Proceedings of the Tenth ACM International Conference on Mobile Computing and Networking (MOBICOM)*, pages 70–84, 2004. 37, 42, 44

[59] M. Ibrahim and M. A. Youssef. Cellsense: A probabilistic rssi-based gsm positioning system. In *Proceedings of IEEE Global Communications Conference (GLOBECOM)*, December 2010. 37, 44

[60] M. A. Yousief. Horus: A wlan-based indoor location determination system. In *A thesis of University of Maryland for the degree of Doctor of Philosophy*, Computer Science Department, University of Maryland, College Park, 2004. 37, 38, 42, 44

[61] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symp. On Math. Stat. and Prob.*, pages 281–296, 1967. 38

[62] P. Bahl and V. N. Padmanabhan. Enhancements to the radar user location and tracking system. *Tech. Rep. MSRTR-2000-12*, Microsoft Research, 2002. 39, 44

[63] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, January 1967. 39, 41

[64] B. Li, J. Salter, A. G. Dempster, and C. Rizos. Indoor positioning techniques based on wireless lan. *Tech. Rep.*, Microsoft Research, 2006. 39, 44

[65] V. Honkavirta, T. Perala, S. Ali-Loytty, and R. Piche. A comparative survey of wlan location fingerprinting methods. In *Proceedings of the 6th Workshop on Positioning, Navigation and Communication (WPNC)*, pages 243–251, 2009. 39, 40, 42

[66] B. D. S. Lakmali and D. Dias. Database correlation for gsm location in outdoor & indoor environments. In *Proceedings of the 4th International Conference on Information and Automation for Sustainability*, pages 42–47, 2008. 39

[67] H. Laitinen, J. Lahteenmaki, and T. Nordstrom. Database correlation method for gsm location. In *Proceedings of the 53rd IEEE Vehicular Technology Conference (VTC2001-Spring)*, pages 2504–2508, 2001. 39

[68] P. Prasithsangaree, P. Krishnamurthy, and P. K. Chrysanthis. On indoor position location with wireless lans. In *Proceedings of IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications*, September 2002. 40

[69] S. Saha, K. Chaudhuri, D. Sanghi, and P. Bhagwat. Location determination of a mobile device using ieee 802.11b access point signals. In *Proceedings of IEEE Wireless Communications and Networking Conference*, March 2003. 40

[70] M. M. Deza and E. Deza. *Dictionary of Distances*. Elsevier Science, 2006. 40

[71] J. T. Tou and R. C. Gonzalez. *Pattern Recognition Principles*. Addison-Wesley, 1974. 40

[72] T. Roos, P. Myllymki, H. Tirri, and P. Misikangas. A probabilistic approach to wlan user location estimation. *International Journal of Wireless Information Networks*, 9(3):155–164, July 2002. 41, 42, 43, 44, 140, 155

[73] K. Kaemarungsi. Distribution of wlan received signal strength indication for indoor location determination. In *Proceedings of the 1st International Symposium on Wireless Pervasive Computing*, pages 6–11, January 2006. 42

[74] V. Honkavirta. Location fingerprinting methods in wireless networks. In *Master thesis*, Tampere University of Technology, 2008. 42

[75] A. Kushki, K. N. Plataniotis, and A. N. Venetsanopoulos. Kernel-based positioning in wireless local area networks. *IEEE Transactions on Mobile Computing*, 6(6):689–705, 2007. 42

[76] A. Kushki, K. N. Plataniotis, A. N. Venetsanopoulos, and C. Regazzoni. Radio map fusion for indoor positioning in wireless local area networks. In *Proceedings of the Eighth International Conference on Information Fusion*, pages 1311–1318, 2005. 42

[77] M. Brunato and R. Battiti. Statistical learning theory for location fingerprinting in wireless lans. *Computer Networks*, 47(6):925–945, April 2005. 43

[78] J. T. Kwok, J. J. Pan, Q. Yang, and Y. Chen. Multidimensional vector regression for accurate and low-cost location estimation in pervasive computing. *IEEE Transactions on Knowledge and Data Engineering*, 18(9):1181–1193, 2006. 43

[79] J. J. Pan, J. T. Kwok, Q. Yang, and Y. Chen. Accurate and low-cost location estimation using kernels. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1366–1371, 2005. 43

[80] P. Chiu, P. Castro, T. Kremenek, and R. Muntz. A probabilistic room location service for wireless networked environments. In *Proceedings of the International Conference on Ubiquitous Computing (Ubicomp)*, pages 18–34, 2001. 43, 44

[81] J. Hightower, D. Fox, L. Liao, D. Schulz, and G. Borriello. Bayesian filtering for location estimation. *IEEE Pervasive Computing*, 2(3):24–33, July 2003. 43

[82] M. A. Youssef, A. Agrawala, A. U. Shankar, and S. H. Noh. A probabilistic clustering-based indoor location determination system. Technical report, Technical Report UMIACS-TR 2002-30 and CS-TR 4350, Computer Science, University of Maryland, March 2002. [Online]. 53

[83] L. Mengual, O. Marban, and S. Eibe. Clustering-based location in wireless networks. *Expert Systems with Applications*, 37(9):6165–6175, 2010. 53

[84] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *SCIENCE*, 315(5814):972–976, February 2007. 53, 54, 59

[85] G. S. V. Vovk and I. Nouretdinov. Self-calibrating probability forecasting. *Advances in Neural Information Processing Systems 16*, 2003. 53, 64

[86] D. Dueck. Affinity propagation: Clustering data by passing messages. In *A thesis of University of Toronto for the degree of Doctor of Philosophy*, Electrical and Computer Engineering, University of Toronto, 2009. 54

[87] Affinity Propagation FAQ. Avaliable:. http://www.psi.toronto.edu/affinitypropagation/faq.html. [Online]. 59

[88] Google Project. location-estimation-trials. http://code.google.com/p/location-estimation-trials/. [Online]. 77, 149

[89] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* 3rd ed.: McGraw-Hill, 2002. 84

[90] Y. Ji. Practical precision bound for indoor location determination. In *Proceedings of International Conference on Computer and Information Application (ICCIA)*, pages 410–413, Tianjin, China, December 2010. 85

[91] B. Silverman. *Density estimation for statistics and data analysis.* Chapman and Hall, 1986. 99

[92] P. Krishnan, A. Krishnakumar, W. H. Ju, C. Mallows, and S. Ganu. A system for lease: Location estimation assisted by stationery emitters for indoor rf wireless networks. In *Proceedings of the 23rd of IEEE 19th Annu. Joint Conf. of the IEEE Comput. and Commun. Soc.*, pages 1001–1011, HongKong, March 2004. 114, 115

[93] A. Haeberlen and A. Rudys. Practical robust localization over largescale 802.11 wireless networks. In *Proceedings of the 10th ACM Int. Conf. on Mobile Computing and Networking*, pages 70–84, New York, 2004. 114, 115

[94] J. Yin, Q. Yang, and L. M. Ni. Learning adaptive temporal radio maps for signal-strength-based location estimation. *IEEE Trans. Mobile Comput.*, 7(7):869–883, July 2008. 114, 115

[95] Weather2. Avaliable:. http://www.myweather2.com/. [Online]. 119

[96] S. ullback, K. P. Burnham, N. F. Laubscher, G. E. Dallal, L. Wilkinson, D. F. Morrison, M. W. Loyer, and B. Eisenberg et al. Letter to the editor: The kullback-leibler distance. *The American Statistician*, 41(4):340-341. JSTOR 2684769, 1987. 132

[97] A. H. S. Ang. and W. H. Tang. *Probability Concepts in Engineering.* John Wiley & Sons, 2007. 134

[98] Jesery. Avaliable:. http://en.wikipedia.org/wiki/Jersey. [Online]. 137

[99] A. Varshavsky, V. Otsason, A. LaMarca, and E. de Lara. Accurate gsm indoor localization. In *Proceedings of the 7th Int. Conf. on Ubiquitous Computing*, pages 141–158, September 2005. 140

[100] O. Vinyals, E. Martin, G. Friedland, and R. Bajcsy. Precise indoor localization using smart phones. In *Proceedings of the 18th International Conf. on Multimedia*, pages 787–790, October 2010. 140

[101] A. S. I. Noh, W. J. Lee, and J. Y. Ye. Comparison of the mechanisms of the zigbee's indoor localization algorithm software engineering. In *Proceedings of the 9th Int. Conf. on Software Eng., Artificial Intell., Networking, and Parallel/Distributed Computing*, pages 13–18, August 2008. 140

[102] H. Abdi. Kendall rank correlation. *Encyclopedia of Measurement and Statistics*, N. J. Salkind, Ed. Thousand Oaks, Calif: SAGE, 2007. 157